Centro de Investigación y de Estudios
Avanzados del Instituto Politécnico Nacional

Unidad Zacatenco

Departamento de Computación

Clasificación por nivel socioeconómico de las regiones geográficas de México

Tesis que presenta

Raúl Maximiliano Urrutia Hernández

Para obtener el Grado de

Maestro en Ciencias

en Computación

Directores de la Tesis:

Dr. Amilcar Meneses Viveros

Dr. Sergio Víctor Chapa Vergara

Ciudad de México **Octubre de 2019**

Resumen

La clasificación es una tarea del aprendizaje automático que permite asignar una categoría a cada elemento de un conjunto de datos, a partir de un conjunto de elementos cuyas categorías son conocidas. De entre todas las aplicaciones que tiene, podemos utilizarla para clasificar a las regiones geográficas del país de acuerdo al nivel socioeconómico de la población de estas. Sin embargo, la información necesaria para determinar el nivel socioeconómico de una región es proporcionada mediante los censos de población y vivienda que efectúa el Instituto Nacional de Estadística, Geografía e Informática (INEGI) con una periodicidad de 10 años.

En el presente trabajo estudiamos la posibilidad de realizar una clasificación de las regiones de México por nivel socioeconómico, con base en el tipo y la cantidad de unidades económicas que estas posean. La fuente de información que utilizamos es el directorio estadístico nacional de unidades económicas (denue), que es un registro de los negocios y establecimientos que existen en el país.

Dado que toda tarea de clasificación requiere de un conjunto de datos clasificados para entrenamiento, primero realizamos un agrupamiento de regiones mediante la información del censo del año 2010 para así tener una aproximación de las clases a las que pertenecen. Posteriormente creamos un conjunto de datos a partir de los datos del denue, con el cual se puede diseñar un modelo de clasificación.

Fueron probados diversos tipos de algoritmos de clasificación; el modelo creado por un bosque aleatorio es el que otorga la mayor exactitud, con un 70 % de regiones correctamente clasificadas para un número de 3 clases.

Abstract

Classification is a machine learning task that assigns a category to each element of a data set, based on a set of elements whose categories are known. Among all the applications it has, we can use it to classify the geographical regions of the country according to the socioeconomic level of its population. However, the information necessary to determine the socioeconomic level of a region is provided through the population and housing censuses carried out by the National Institute of Statistics, Geography and Informatics (INEGI) with a periodicity of 10 years.

In this work we study the possibility of making a classification of the regions of Mexico by socioeconomic level, based on the type and quantity of economic units in them. The source of information we used is the national statistical directory of economic units (denue), which is a record of the businesses and establishments that exist in the country.

Since every classification task requires a set of classified data for training, we first performed a clustering of the regions using the information from the census of the year 2010 in order to have an approximation of the classes to which they belong. Subsequently we created a data set from the denue data, with which a classification model can be designed.

Various types of classification algorithms were tested; the model created by a random forest is the one the grants the highest accuracy, with 70 % of the regions correctly classified for 3 classes.

Agradecimientos

Agradezco al CINVESTAV por brindarme la formación y el apoyo para realizar mis estudios de maestría y al CONACYT por proporcionarme el apoyo económico para sustentar mi estancia durante estos dos años de estudio.

A mis asesores, el Dr. Amilcar Meneses y el Dr. Sergio Chapa, por recibirme como su estudiante y brindarme su confianza, paciencia y toda la ayuda que necesité para la elaboración de este trabajo.

A los revisores de esta tesis, la Dr. Sonia Mendoza y la M. en C. Erika Hernández, por tomarse el tiempo para revisar esta tesis y por sus observaciones.

A mis padres, Raúl e Inés, por su incondicional e inmensurable apoyo, sus consejos y su cariño. Sin ellos este logro no hubiera sido posible.

A Fernanda, mi novia, porque a pesar de la distancia siempre ha estado dispuesta a brindarme todo su apoyo, amor y comprensión.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En México, el INEGI realiza censos de población y vivienda en periodos de 10 años, en los cuales presenta conteos y datos estadísticos de las principales características sociales, económicas y culturales de la población residente del país. Con esta información pueden construirse indicadores que nos permitan estudiar a la población y ayudarnos a identificar algunos problemas que esta pueda tener. Uno de los indicadores sociales más importantes es el nivel socioeconómico (NSE). Tanto el gobierno, el sector social, el académico y privado, pueden utilizarlo como apoyo para estrategias y programas específicos de acuerdo a su ámbito de acción; por ejemplo, la asignación de recursos de inversión para el desarrollo social de una zona, o para el expasión de una empresa.

Con la Ciencia de Datos, podemos procesar datos como el del censo población y vivienda para extraer conocimientos de estos, o bien mejorar su interpretación. También se puede hacer uso de algunas herramientas de aprendizaje automático para hacer predicciones acerca de las condiciones de la población respecto a algún indicador socioeconómico.

1.1. Nivel Socioeconómico

El nivel socioeconómico es un indicador que comprende un conjunto de variables económicas, sociológicas, educativas y laborales por las que se califica a un individuo o un colectivo.
La medición del nivel socioeconómico no está estandarizada, los criterios pueden variar de país a país, además de que con el paso del tiempo se pueden incorporar nuevos factores o desechar algunos.

En el año 2004, el INEGI, en su proyecto "Regiones socioeconómicas de México"[INEGI, 2004], realizó una clasificación de las regiones del país de acuerdo a sus características socioeconómicas. Utilizó indicadores agrupados de la siguiente manera:

a) Infraestructura de la vivienda.

b) Calidad de la vivienda.

c) Hacinamiento.

d) Equipamiento en la vivienda.

e) Salud.

f) Eduación.

g) Empleo.

En este producto el INEGI dividió el nivel socioeconómico en 7 estratos distintos a nivel estatal, municipal y de área geoestadística básica (ageb). Todos los indicadores fueron construidos a partir de la información recabada en el censo de población y vivienda del año 2000.

Otra institución reconocida en el país que realiza estudios socioeconómicos, es la Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión Pública (AMAI), la cual aplica un modelo estadístico que permite agrupar y clasificar a los hogares dentro de 7 estratos [AMAI, 2018]. Para calcular las distribuciones, la AMAI utiliza información recopilada por distintas encuentas realizadas por INEGI, principalmente la Encuesta Nacional de Ingresos y Gastos de los Hogares (enigh).

El modelo que utiliza AMAI considera seis características a nivel de hogares:

a) Escolaridad del jefe del hogar.

b) Número de dormitorios.

c) Número de baños completos.

d) Número de personas ocupadas de 14 años y más.

e) Número de autos.

f) Tenencia de internet.

La estratificación la realizan mediante la regla AMAI 8x7, la cual utiliza un sistema de puntos para cada variable y sus distintos niveles. Con los puntajes se obtiene una suma de puntos para todos los hogares y posteriormente se utiliza el procedimiento de estratificación univariado de Dalenius-Hodges para obtener los puntos de corte que minimizan la variabilidad intragrupos.

1.2. Aprendizaje Automático

El aprendizaje automático es una rama de la inteligencia artificial que estudia algoritmos y modelos estadísticos para que los sistemas computacionales puedan realizar distintas tareas sin utilizar instrucciones específicas, sino patrones e inferencias. A menudo es definido como el proceso automático para extraer patrones a partir de datos.

Dentro del aprendizaje automático supervisado, los algoritmos contruyen un modelo matemático de la relación entre un conjunto de características descriptivas y una característica objetivo, basándose en un conjunto de ejemplos o instancias. Este modelo se usa posteriormente para realizar predicciones para nuevas instancias. Dos tipos muy conocidos de aprendizaje automático supervisado son la clasificación y la regresión. Los algoritmos de clasificación se utilizan cuando las características de salida están restringidas a un conjunto limitado o discretizado, mientras que los algoritmos de regresión se utilizan cuando se presentan características de salida continuas.

Otra categoría del aprendizaje automático es el aprendizaje no supervisado, en el cual los algoritmos realizan un modelo a partir de un conjunto de datos que tiene características de entrada, pero ninguna característica deseada o de salida. Los algoritmos de este tipo se utilizan para encontrar estructuras en los datos, como un agrupamiento.

1.3. Descripción del Problema

Actualmente ya se cuenta con información concerniente al estrato socioeconómico de las regiones del país. Sin embargo, las instituciones que lo determinan usan distintas fuentes de información, distintos indicadores y diferentes ventanas de tiempo, por lo que la información puede no ser equivalente entre estas instituciones y puede además no estar actualizada. El proyecto "Regiones socioeconómicas de México"de INEGI se realizó en el año 2004 con base en información del censo del 2000, por lo que en la actualidad dicha información no es muy confiable, puesto que muchas regiones probablemente han mejorado (o empeorado) su nivel social y económico en el transcurso de estas casi dos décadas. INEGI no realizó el mismo estudio para el censo del 2010, que es el más reciente; aún así, existen nueve años de diferencia a fecha de hoy. El uso de información más actualizada puede permitirnos realizar un estudio socieconómico de las regiones del país que refleje de mejor manera la actual situación de las mismas. La encuesta enigh que utiliza AMAI en su estudio tiene una periodicidad de dos años, por lo que sin duda alguna presenta resultados más actualizados utilizando dicha información, que la del censo. No obstante, a diferencia del censo, dicha encuesta tiene el inconveniente de que no se realiza en todos los hogares del país, sino que toma una muestra que representa a toda una región.

En nuestro trabajo proponemos el uso de herramientas de aprendizaje automático para construir un modelo de clasificación a nivel de área geoestadística básica, de acuerdo al nivel socioeconómico y por medio de los establecimientos y negocios que hay en ellas. Los datos con los que construimos el modelo provienen del directorio estadístico nacional de unidades económicas (denue). En una primera instancia, realizamos un agrupamiento de las ageb para poder etiquetarlas y así tener un conjunto de entrenamiento para el modelo. Posteriormente se prueban varios algoritmos de clasificación para seleccionar el que construya el modelo con la mejor exactitud.

1.4. Objetivos

1.4.1. Objetivo general

Crear un modelo que determine el nivel socioeconómico al que pertenecen las ageb del país, con la información de las unidades económicas que hay en ellas.

1.4.2. Objetivos específicos

- Construir un conjunto de datos con los indicadores de nivel socioeconómico que puedan obtenerse de la información del censo de población y vivienda 2010.
- Realizar un agrupamiento de las ageb para etiquetarlas y así poder construir un conjunto de datos de entrenamiento.
- Analizar la distribución de los tipos de unidades económicas que existen para así conocer la tendencia con la que estas están presentes en las agebs de distintos niveles socioeconómicos.
- Construir un conjunto de datos que contenga información sobre el tipo y el número de unidades económicas que posee cada ageb.
- Probar diversos algoritmos de clasificación para clasificar a las ageb por nivel socioeconómico.
- Interpretar resultados.

1.5. Estructura del documento

El resto del presente documento se encuentra organizado de la siguiente manera: Los capítulos 2 y 3 constituyen el marco teórico de esta tesis. En el primero se abordan temas de análisis de datos, que incluyen conceptos sobre conjuntos de datos y los procedimientos que se aplican en ellos como paso previo al aplicar algoritmos de aprendizaje automático. En el capítulo 3 se exponen descripciones sobre los métodos de agrupamiento y clasificación que utilizamos en nuestro trabajo. En el capítulo 4 se explica el concepto de nivel socioeconómico, los indicadores comúnmente empleados para determinarlo y algunos trabajos que se han realizado al respecto. El capítulo 5 describe el método que empleamos para el agrupamiento y para la construcción del modelo de clasificación, así como los conjuntos de datos empleados. En el capítulo 6 se dan a conocer los resultados de las pruebas con los algoritmos de clasificación. Por último, en el capítulo 7 se mencionan nuestras concluciones y una pequeña discusión del trabajo a futuro.

Capítulo 2

Análisis de Datos

En este capítulo se presentan algunos tópicos relacionados a conjuntos de datos que son importantes para nuestro trabajo. La mayoría de la información aquí presente fue tomada de [Tan et al., 2013].

2.1. Estructura y tipos de Datos

La estructura de los datos es la manera en la que la información está organizada y almacenada, de modo que pueda ser accedida y modificada eficientemente. Más precisamente, una estructura de datos es una colección de valores, relaciones y funciones que podemos aplicar sobre ellos. Existen tres tipos de estructuras en las que podemos encontrar la información:

Datos estructurados: Los elementos de datos estructurados están sujetos a un modelo predefinido que los hace fáciles de analizar. Se conforma forma tabular con relaciones entre las diferentes filas y columnas. Algunos ejemplos de datos estructurados son archivos de Excel o bases de datos SQL.

Datos no estructurados: Los datos no estructurados no están organizados de manera predefinida ni sujetos a un modelo de datos. Estos datos se pueden encontrar como archivos de texto o archivos en formato PDF, entre otros.

Datos semiestructurados: La información semiestructurada no se ajusta a la estructura de bases de datos o tablas, pero sí contiene etiquetas o marcadores que permite separar elementos y hacerlos más fáciles de analiozar. Ejemplos de datos semiestructurados son archivos XML y JSON.

En adelante utilizaremos únicamente conjuntos de datos estructurados.

Un conjunto de datos es una colección de objetos (también llamados registros, entradas, eventos u observaciones) que están descritos por un número de atributos (o variables) que capturan sus características básicas. Un atributo es una propiedad que puede variar de un objeto a otro, por ejemplo, la estatura y el color de ojos varía de

una persona a otra. Sin embargo, podemos notar que el color de los ojos es un atributo simbólico con un determinado número de valores posibles (cafés, negros, azules, verdes, etc.), mientras que la estatura es un atributo numérico con un amplio rango de posibles valores. Es común dividir a los tipos de atributos en categóricos y numéricos, los cuáles también se pueden subdividir de acuerdo la figura 2.1.

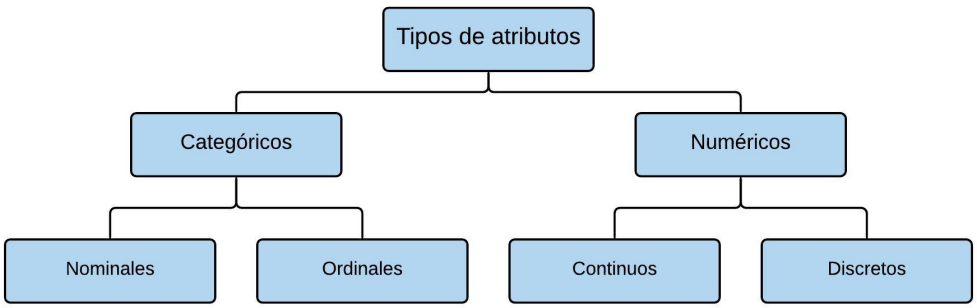

Figura 2.1: Diagrama de los tipos de atributos.

Los atributos nominales proveen información que permita distinguir a los objetos; algunos ejemplos pueden ser el color de ojos, números de identificación o matrícula, códigos postales, etc. Los atributos ordinales, como los resultados de una encuesta de satisfacción, permiten ordernar a los objetos, aunque la diferencia entre cada uno de ellos pueda ser desconocida. Alternativamente, los atributos númericos se pueden dividir en atributos de intervalo y de proporción. Los primeros son escalas numéricas en las que se conocen el orden y las diferencias exactas entre los valores, el ejemplo clásico de este tipo de atributo es la temperatura en grado Celsius o Fahrenheit. En los atributos de intervalo no se pueden establecer proporciones, ya que, por ejemplo, $20°C$ no es el doble de caliente que $10°C$, puesto que al convertir a grados Fahrenheit obtenemos $10°C = 50°F$ y $20°C = 68°F$. Por otra parte los atributos de proporción, como la temperatura en Kelvin, proveen información del valor exacto entre unidades y poseen un zero absoluto.

Es importante saber diferenciar a los tipos de atributos ya que permiten establecer qué operaciones podemos aplicar en ellos para su análisis.

2.2. Calidad de los Datos

Muchos de los datos en crudo obtenidos de conjuntos de datos se encuentran incompletos, con ruido o sin procesar. Por ejemplo, los conjuntos de datos pueden contener:

Ruido: Puede involucrar la distorsión de un valor o la presencia de objetos distorsionados.

Valores faltantes: Es posible que algunos objetos tengan valores faltantes en uno o más atributos, ya sea porque la información no fue proporcionada o porque algunos atributos pueden no ser aplicables para todos los objetos.

Valores atípicos: Incluyen objetos que tienen características muy distintas a la mayoría del resto de los objetos del conjunto de datos o valores de algún atributo poco comunes respecto a los valores típicos del mismo.

Valores inconsistentes: Generalmente debidos a datos mal proporcionados o errores de captura. Algunas inconsistencias son fáciles de detectar.

Datos duplicados: Un conjunto de datos puede incluir objetos duplicados, o incluso dos objetos que representen a uno mismo aunque difieran en alguno de sus atributos.

Existen distintas maneras de lidiar con valores faltantes, cada una siendo apropiada dependiendo de las circunstancias. La más utilizada y simple es eliminar los objetos o atributos que tengan valores faltantes. Es recomendable eliminar objetos cuando son sólo una pequeña porción del conjunto los que presentan estos valores, mientras que la eliminación de atributos debe hacerse con el cuidado de no eliminar algún atributo que sea importante para el análisis. Alternativamente se pueden reemplazar los valores faltantes por alguna constante generada a partir de la distribución de la variable o bien, a partir de las demás características del objeto; también se puede optar por simplemente ignorarlos.

Al tratar con valores duplicados, se debe tener precaución en no combinar objetos que son similares (como dos personas con el mismo nombre) y en no confundir objetos que sean distintos pero que tengan características idénticas.

Es importante lidiar con estas cualidades para realizar un análisis de los datos más apropiado. A pesar de que la mayoría de los algoritmos de minería de datos y aprendizaje automático pueden tolerar cierto nivel de imperfección en los datos, una mejora en la calidad de los datos usualmente implica una mejora del análisis final.

2.3. Preprocesamiento de Datos

Procesar los datos en crudo permite, además de mejorar la calidad de los datos, modificarlos para que se ajuste de mejor manera a alguna técnica o algoritmo en específico. Algunos enfoques que podemos emplear son:

- Agregación.
- Muestreo.
- Reducción de la dimensionalidad.
- Selección de variables.

- Creación de variables.
- Transformación de variables.

En términos generales, se pueden seleccionar objetos y atributos para su análisis, o bien, crear o cambiar atributos; todo con el objetivo de hacer más eficiente el análisis, ya sea en términos de tiempo, costo o calidad. A continuación explicamos en qué consisten algunos de estos enfoques.

2.3.1. Agregación

Es cualquier proceso en el cual la información de un conjunto de datos es reunida y expresada de una forma resumida. La agregación nos puede ayudar a reducir la memoria requerida y el tiempo de procesamiento, lo que permite emplear algoritmos que puedan ser costosos. También nos puede ayudar a obtener más información respecto a grupos particulares de acuerdo a características específicas.

Una desventaja de la agregación es la pérdida de interpretación y de detalles interesantes.

2.3.2. Muestreo

El muestreo es una técnica utilizada para seleccionar un subconjunto de objetos para ser analizado. Esto es muy útil cuando el conjunto de datos inicial es muy grande, lo que hace que procesar los datos sea costoso.

La muestra seleccionada debe ser representativa del conjunto de datos, es decir, que tenga aproximadamente las misma propiedades de interés. El muestreo aleatorio es el método más simple de muestreo, en el cual cada objeto tiene la misma probabilidad de ser seleccionado; este muestreo se puede realizar con reemplazo y sin reemplazo. Sin embargo, para poblaciones compuestas por objetos de distintos tipos o clases con un números muy distintos de objetos, el muestreo aleatorio puede fallar en seleccionar muestras representativas de las clases menos frecuentes. Ante esta situación, se puede optar por un muestreo estratificado, en el que se inicia con un grupo predefinido de objetos y a partir de este, seleccionar nuevos objetos.

2.3.3. Reducción de la dimensionalidad

La dimensionalidad de un conjunto de datos es el número de atributos que poseen los objetos. La mayoría de los algoritmos de aprendizaje automático trabajan mejor si la dimensionalidad es baja. Además, reducir la dimensionalidad puede eliminar atributos irrelevantes, reducir el ruido y permitir una mejor visualización de los datos. Aunque la dimensionalidad no se reduzca a dos o tres dimensiones, normalmente los datos se visualizan mediante pares o tripletas de atributos. También se reduce la memoria requerida y el tiempo de ejecución de los algoritmos.

Emplear reducción de la dimensionalidad también puede evitar los fenómenos de maldición de la dimensionalidad. Este conjunto de fenómenos se presentan cuando el análisis de datos se dificulta significativamente conforme la dimensionalidad aumenta, ya que el volumen del espacio aumenta tan rápido que ocasiona que los datos se dispersen. Para clasificaciones, esto puede significar que no hay suficientes objetos que permitan la creación de un modelo que asigne correctamente una clase a todos los objetos, mientras que para agrupamiento, los conceptos de densidades y distancias entre puntos del espacio pierden significado. Estos problemas resultan en una exactitud deficiente para clasificación, o grupos de baja calidad para agrupamiento.

Las técnicas más empleadas para reducción de la dimensionalidad son técnicas de álgebra lineal que mapean los datos de un espacio de gran dimensión a uno de menor dimensión. Destacan el *análisis de componentes principales* (PCA, por sus siglas en inglés), la *descomposición en valores singulares* (SVD, por sus siglas en inglés) y el *análisis factorial*. En la siguiente sección se explican las ideas básicas de PCA.

2.3.4. Análisis de componentes principales

El procedimiento PCA consiste en encontrar nuevos atributos (componentes principales) que sean combinación lineal de los atributos originales, que sean ortogonales entre sí, y que representen a la mayor cantidad de varianza original. Las nuevas componentes se ordenan conforme a la cantidad de varianza que describen. Para explicar el funcionamiento de PCA, primero debemos definir la covarianza de dos atributos.

Definición 2.3.1. Dada una matriz de datos D de m filas (objetos) y n columnas (atributos), la covarianza de dos atributos, i y j, de D, se define como

$$Cov(i,j) = \frac{1}{n}\sum_{k=1}^{m} d'_{ki}d'_{kj},$$

donde d'_{ki} denota la desviación de la componente k-ésima del atributo i respecto a la media, es decir $d'_{ki} = d_{ki} - \mu_i$. La covarianza de dos atributos es un medida de la tendencia en la que varían uno con respecto al otro. Si $i = j$, es decir, los atributos son los mismos, la covarianza pasa a ser la varianza del atributo.

Definición 2.3.2. La matriz S de $n \times n$ con componentes $s_{i,j} = Cov(i,j)$, se conoce como matriz de covarianza de D.

Al pedir como requisito que todos los datos estén centrados a una media de cero ($\mu_i = 0\,\forall i \in [1,n]$), tendremos que $S = D^T D$. La matriz de covarianza es una matriz semidefinida positiva, lo que hace que tenga, entre otras propiedades, valores propios mayores o iguales a cero. Esto nos permite ordenar a los valores propios, $\lambda_1, \ldots, \lambda_n$, de S. Si $U = [u_1, \ldots, u_n]$ es la matriz de vectores propios ordenados de S, entonces el i-ésimo vector propio corresponde al i-ésimo valor propio más grande. Con esto en mente, podemos establecer lo siguiente.

- La matriz de datos $D' = DU$ tiene la propiedad de que, la covarianza de cada par de atributos distintos es igual a cero, además de que los datos están ordenados respecto a su contribución a la varianza total.

- Cada nuevo atributo es una combinación lineal de los atributos originales, donde los pesos de la combinación lineal para el i-ésimo atributo son las componentes del i-ésimo vector propio.

- La varianza del i-ésimo nuevo atributo es λ_i.

- La suma de las varianzas de los nuevos atributos es igual a la suma de las varianzas de los atributos originales.

Los nuevos atributos de D' se conocen como *componentes principales*. El vector propio asociado con el valor propio más grande, indica la dirección en la cual los datos presentan una mayor varianza, mientras que el vector propio asociado al segundo valor propio más grande indica la dirección en la que los datos presentan la mayor porción de la varianza restante; y así sucesivamente. Es así como funciona PCA. Es importante escalar las variables, ya que de lo contrario, la primer componente principal tendrá una carga muy grande hacia el atributo que tenga la mayor varianza [James et al., 2013].

En general, una matriz de datos D de $m \times n$ tiene mín$(m-1, n)$ componentes principales distintas, sin embargo, casi nunca estamos interesados en todas ellas; de hecho, cuando se quiere visualizar la información o intepretar los datos, se suele utilizar sólo el primer par de componentes pricipales. Lo que se busca usualmente es utilizar la menor cantidad de componentes principales requeridas para un buen entendimiento de los datos. Dependiendo del problema que se tenga, uno toma el número de componentes que expliquen la varianza que se necesite. Para problemas de análisis supervisado, el número de componentes principales indicado puede obtenerse al ser usado como parámetro de ajuste.

En la tabla 2.1 se describe la varianza explicada por las componentes principales del conjunto de datos Iris. Este conjunto de datos es muy utilizado para pruebas y ejemplos de algoritmos estadísticos y de aprendizaje automático; contiene 50 muestras de las tres especies de la flor iris (setosa, virginica y versicolor) con cuatro atributos para cada muestra: el largo y el ancho del pétalo y del sépalo.

Componente Principal	Proporción de varianza explicada
1	0.7296
2	0.2285
3	0.03669
4	0.00518

Tabla 2.1: Proporción de la varianza explicada por las componentes principales del conjunto de datos Iris

De las cuatro componentes principales, puede observarse que la tercer y cuarta componentes contribuyen con una baja proporción a la varianza respecto a las primeras dos, en especial la cuarta, con sólo un 0,00518. En la figura 2.2 se pueden apreciar gráficamente las contribuciones a la varianza.

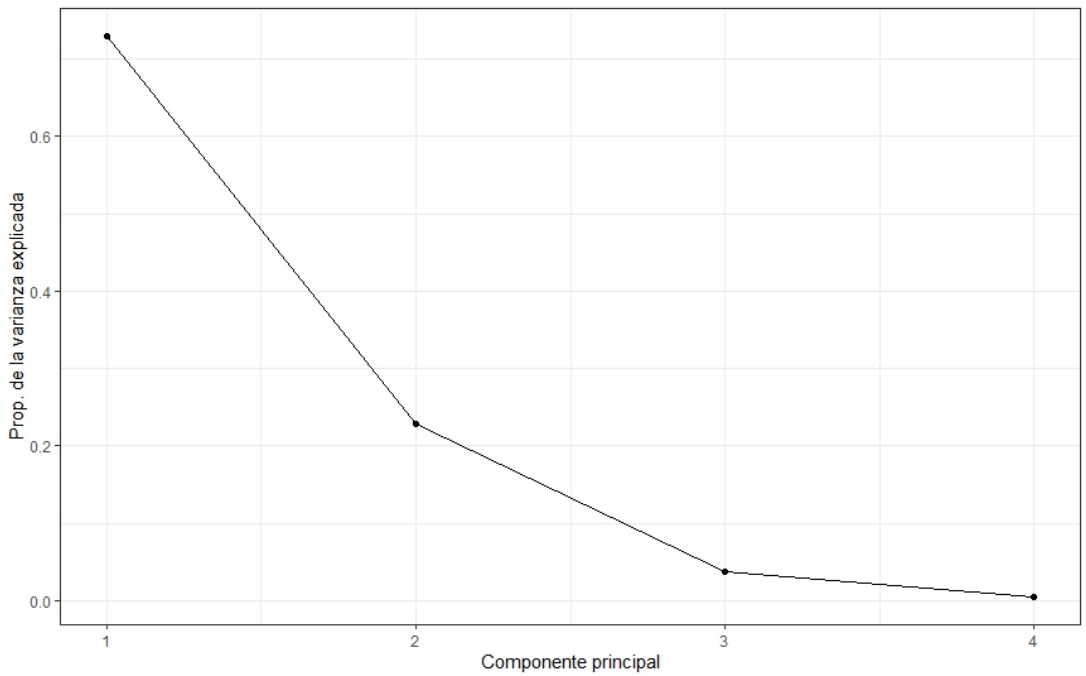

Figura 2.2: Gráfico de la proporción de la varianza explicada por cada una de las componentes principales del conjunto de datos Iris

La varianza acumulada explicada por las componentes principales está graficada en la figura 2.3. Podemos observar que las dos primeras componentes pincipales representan la mayor parte de la varianza total (0,9581), por lo que podríamos trabajar únicamente con ellas para análisis posteriores.

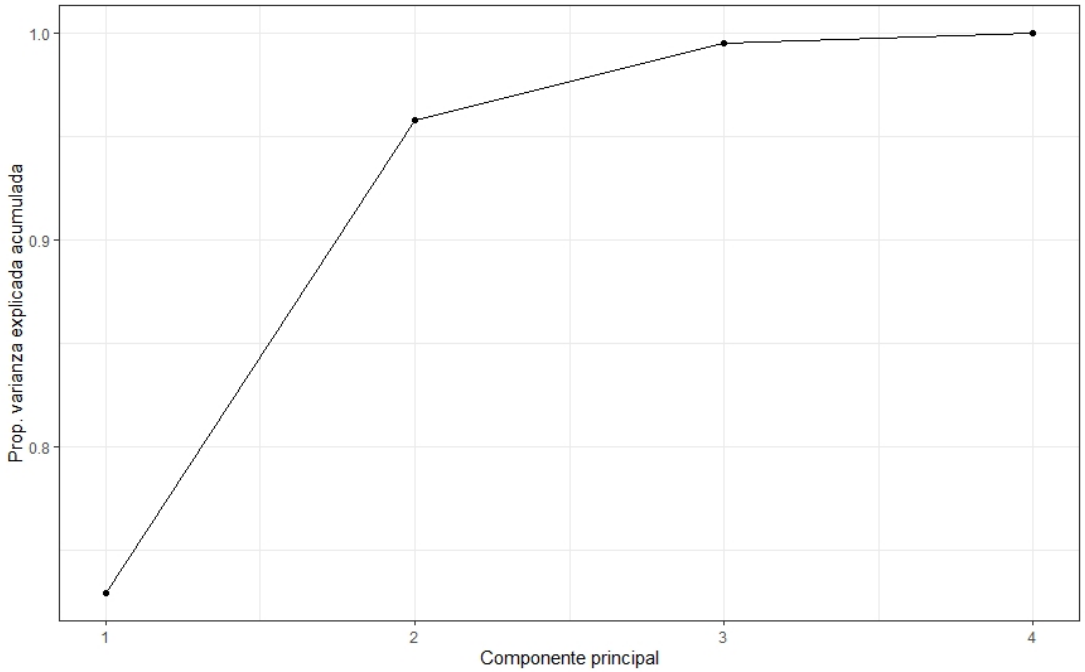

Figura 2.3: Gráfico de la proporción de la varianza acumulada explicada por las componentes principales del conjunto de datos Iris

2.3.5. Selección de variables

Las estrategias de reducción de la dimensionalidad generalmente están diseñadas para crear nuevas variables que son combinaciones de las originales. Por otra parte, la selección de variables consiste en escoger un subconjunto de los atributos originales. Aunque eliminar variables parezca que llevará a la pérdida de información, no es así cuando se trata de variables poco relevantes o redundantes; por ejemplo, la taza de población alfabeta y la de población analfabeta son dos atributos redundantes, ya que básicamente explican la misma información. Este tipo de variables pueden afectar negativamente la exactitud en problemas de clasificación o la calidad de los grupos en problemas de agrupamiento.

El método ideal para seleccionar un subconjunto de variables adecuado, sería evaluar todos los posibles subconjuntos de atributos en el algoritmo que se vaya a utilizar y seleccionar aquel que produzca los mejores resultados. Desafortunadamente, dicho método es poco práctico en la mayoría de situaciones, puesto que existen 2^n subconjuntos posibles de un conjunto de n variables. Existen tres enfoques a los que podemos recurrir:

Integrados: Aquí la selección de variables se realiza como parte del proceso de contrucción del modelo en algunos algoritmos, los cuales deciden por sí mismos cuáles atributos utilizar y cuáles ignorar. Los clasificadores con árboles de decisión son algoritmos de este tipo.

Envolvedores: Este tipo de métodos utilizan el modelo predictivo final a emplear como caja negra para encontrar el mejor subjunto de atributos, de manera similar al método ideal, pero sin enumerar a todos los posibles subconjuntos. Los métodos envolvedores entrenan un nuevo modelo para cada subconjunto, lo que es computacionalmente intensivo, aunque usualmente entrega el mejor subconjunto que se ajuste al modelo.

Filtrado: Las variables se seleccionan antes de aplicar los algoritmos, empleando alguna medida independiente de estos. Entre las medidas más comunes están el coeficiente de correlación de Pearson, el punto de información mutua y la distancia intragrupos. Estos métodos suelen ser menos intensivos que los envolvedores, aunque producen un subconjunto que no está apegado a ningún modelo predictivo.

2.3.6. Creación de variables

Con frecuencia es posible crear un conjunto nuevo de atributos a partir de los originales, que permitan capturar la información de mejor manera. Incluso puede suceder que este conjunto sea menor que el original y así también obtener los beneficios ya mencionados de reducir la dimensionalidad. Los procesos de creación pueden ser por extracción (crear un nuevo conjunto de datos a partir de la información en crudo original), por mapeo de datos a un nuevo espacio (mediante transformaciones como la trasformada de Fourier) o por construcción de nuevos atributos que puedan ser más útiles (ya que a veces no se puede aprovechar directamente a los datos originales).

2.3.7. Transformación de variables

Los atributos tienden a tener rangos que pueden variar significativamente unos de los otros. Por ejemplo, una variable puede tomar valores en un rango de 0 a 1, mientras que los de otra variable de 0 a 100. Para algunos algoritmos, estas diferencias pueden ocasionar una tendencia excesiva hacia la variable que tenga un mayor rango. También pueden haber casos en los que únicamente importe la magnitud de una variable, para lo cual la variable se transformar tomando el valor absoluto. A veces también se suelen aplicar funciones simples sobre los atributos para, por ejemplo transoformar los datos a una distribución gaussiana.

Para reducir los efectos de las diferencias de rangos entre las variables, se suelen normalizar o estandarizar. Algoritmos como los de las *redes neuronales artificiales* suelen beneficiarse de datos normalizados, al igual que otros que dependan de medidas en las distancias, como el algoritmo de los *k vecinos más cercanos*. A continuación examinaremos dos de los métodos de normalización más utilizados.

Normalización Min-Max

Esta normalización se basa en observar qué tan grandes son los valores de una variable respecto al menor, para entonces reescalar esta diferencia mediante el rango. Esto es

$$x_{mm} = \frac{x - \text{mín}(x)}{range(x)} = \frac{x - \text{mín}(x)}{\text{máx}(x) - \text{mín}(x)}. \tag{2.1}$$

Los nuevos atributos estarán entonces en un rango del 0 al 1 [Larose and Larose, 2014].

Normalización por puntuación estándar

También conocida simplemente como estandarización (o en inglés Z-score), funciona tomando la diferencia entre el valor del atributo y el valor medio del atributo, y luego dividiendo esta diferencia por la desviación estándar del atributo, es decir

$$x_z = \frac{x - \mu(x)}{\sigma(x)}. \tag{2.2}$$

Esto crea una variable con media igual a 0 y desviación estándar 1 [Larose and Larose, 2014].

2.4. Correlación entre variables

La correlación entre dos variables, x y y, de un conjunto de datos es una medida de la relación lineal entre estas. Existen diferentes tipos de coeficientes de correlación, el más común es el de Pearson, que se define como

$$\rho_{xy} = \frac{Cov(x, y)}{\sigma(x)\sigma(y)}. \tag{2.3}$$

Puede probarse, mediante la desigualdad de Cauchy-Schwarz, que $\rho_{xy}^2 \leq 1$ [Downey, 2011], por lo que $-1 \leq \rho_{xy} \leq 1$. La magnitud de ρ_{xy} indica qué tan fuertemente están correlacionadas x y y; si $\rho_{xy} = 1$ las variables están perfectamente correlacionadas, lo que significa que conociendo una, la otra puede predecirse. Algo similar ocurre si $\rho_{xy} = -1$, pues significa que están negativamente correlacionadas, pero para propósitos de predicción, una correlación negativa es igual de buena que una positiva. Sin embargo, una correlación de 0 no necesariamente implica que no hay relación alguna entre las dos variables, sino simplemente que no tienen una relación lineal.

2.5. Detección de Valores Atípicos

Los valores atípicos (o anomalías) son valores distantes que van en contra de la tendencia del resto de los datos. Identificarlos es importante, ya que pueden representar errores

en los conjuntos de datos; incluso si los valores atípicos son entradas válidas, algunos algoritmos y métodos estadísticos son sensibles a la presencia de estos, con lo que la calidad de los resultados puede verse afectada negativamente.

Los métodos de detección de valores atípicos se pueden dividir en tres categorías dependiendo de la disponibilidad de valores atípicos conocidos:

Detección supervisada Las técnicas de este tipo requieren de la existencia de un conjunto de datos de entrenamiento que contenga objetos de dos clases: objetos normales y anómalos

Detección no supervisada Cuando no es posible etiquetar a los objetos, se asigna un criterio para determinar si un objeto es anómalo. Para que este tipo de técnicas funcionen correctamente, las anomalías deben ser distintas entre ellas, puesto que si hay muchas anomalías que sean similiares, se pueden llegar a identificar como valores normales.

Detección semisupervisada En este tipo de enfoque tiene como objetivo el encontrar un criterio de anomalía para un conjunto de objetos, utilizando la información de objetos etiquetados como normales.

La mayoría de las técnicas para detectar valores atípicos pueden dividirse en tres enfoques de acuerdo a sus definiciones de valor atípico: técnicas estadísticas, técnicas basadas en proximidad y técnicas basadas en densidad. En determinados casos es posible identificar a los valores atípicos de manera gráfica, como se puede observar en la figura 2.4

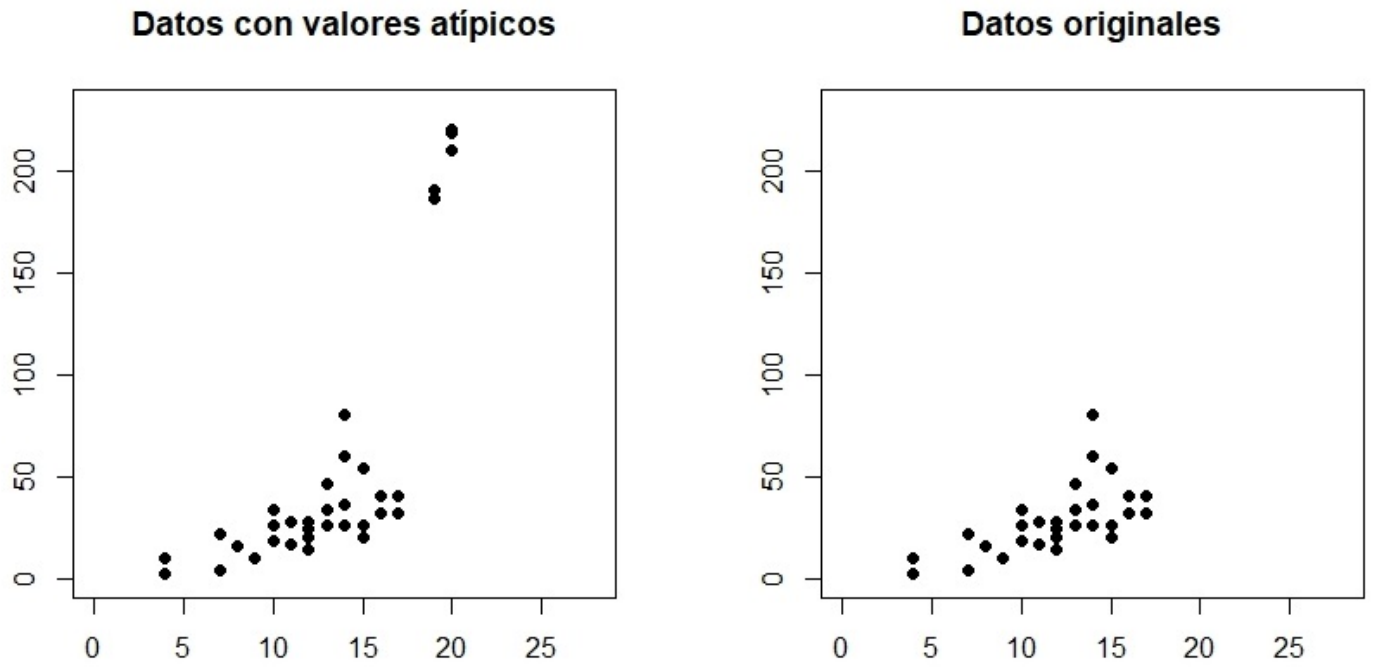

Figura 2.4: Gráficas de dispersión de datos ejemplo con valores atípicos (izquierda) y datos normales (derecha).

2.5.1. Técnicas estadísticas

Este tipo de técnicas contruyen un modelo para los datos que evalúan a los objetos de acuerdo a qué tan bien se ajusten a dicho modelo. En su mayoría, las técnicas estadísticas construyen un modelo de distribución de probabilidad teniendo en cuenta la probabilidad de que los objetos estén bajo ese modelo. La definición probabilista de valor atípico es la siguiente:

Definición 2.5.1. Un valor atípico es un objeto que tiene una baja probabilidad respecto a un modelo de distribución de probabilidad de los datos.

A pesar de que muchos tipos de datos puedan describirse con distribuciones comunes como la gaussiana, la binomial o la de Poisson, es muy frecuente encontrar conjuntos de datos sin ninguna distribución estándar, y si se elige el modelo incorrecto, muchos objetos pueden ser identificados erroneamente como atípicos. Aunque la mayor parte de las técnicas estadísticas aplican para una sola variable, algunas también se han definido para varias variables (aunque para dimensiones grandes, suelen tener un desempeño pobre). Tal es el caso de la distribución normal, en la que, para más de una variable se emplea la distancia de Mahalanobis como medida de la forma en la que están distribuidos los datos. La distancia de Mahalanobis entre un punto x y la media de los datos μ_x está dada por

$$Mahalanobis(x, \mu_x) = (x - \mu_x)S^{-1}(x - \mu_x)^T,$$

donde S es la matriz de covarianza de los datos.

2.5.2. Técnicas basadas en proximidad

La idea básica en estas técnicas es que un objeto es considerado anómalo si está distante de la mayoría de los datos. Este enfoque es más general y más sencillo de aplicar que los métodos estadísticos, puesto que es más fácil determinar una medida de proximidad para un conjunto de datos, que determinar su distribución estadística. Una de las maneras más simples de medir si un objeto está distante de la mayoría es estableciendo un puntaje de acuerdo a la distancia a su k vécino más cercano. Con esto se puede establecer la siguiente definición.

Definición 2.5.2. El puntaje atípico de un objeto está dado por la distancia a su k vecino más cercano.

El puntaje atípico puede ser muy sensible al valor de k, de manera que, si k es muy pequeño, entonces un número pequeño de anomalías cercanas puede causar un alto puntaje atípico.

A pesar de que estas técnicas sean sencillas, suelen tener una complejidad $O(m^2)$, con m el número de objetos del conjunto de datos. Además, son sensibles a la elección de los parámetros y no pueden lidiar con regiones de distintas densidades.

2.5.3. Técnicas basadas en densidad

Desde el punto de vista de densidades, los valores atípicos son objetos que pertenecen a regiones de baja densidad.

Definición 2.5.3. El puntaje atípico de un objeto está dado por el inverso de la densidad alrededor de un objeto.

La densidad alrededor de un objeto puede ser descrita como el número de objetos que están dentro de una distancia d específica del objeto. El parámetro d debe ser elegido apropiadamente, pues para valores muy pequeños, muchos objetos normales podrían tener una baja densidad y entonces un alto puntaje atípico. Si por el contrario d es muy grande, puede que muchos objetos atípicos tengan densidades muy similares a las de objetos normales.

Al igual que las ténicas basadas en proximidad, la mayoría de las técnicas basadas en densidad también tienen complejidad $O(m^2)$, sin embargo, algunas han sido desarrolladas para tratar regiones de distintas densidades y lidiar mejor con la selección de parámetros, entre los que podemos destacar el LOF (Local Outlier Factor) [Breunig et al., 2000] y el algoritmo RKOF (Robust Kernel-based Outlier Factor) [Gao et al., 2011]. Este último tiene además el plus de lidiar con conjuntos de datos grandes con una complejidad $O(m \log(m) + nk)$ (k es un parámetro que determina la escala de la vecindad local).

2.5.4. Técnicas basadas en agrupamiento

Como veremos en el capítulo siguiente, el análisis de grupos consiste en encontrar grupos de objetos fuertemente relacionados. La idea de utilizar análisis de grupos para detección de valores atípicos es natural, debido a que los objetos anómalos no están fuertemente relacionados a ningún otro objeto. Una manera de emplear estas técnicas es eliminando los grupos pequeños que se encuentran alejados de los demás grupos, aunque esto requiere la utilización de parámetros como el tamañó mínimo del grupo y la distancia entre un grupo pequeño con otros grupos. Otra forma de proceder, es primero agrupar todos los objetos del conjunto de datos para después evaluar el grado en que un objeto pertenece a un grupo.

Definición 2.5.4. Un objeto es atípico (con base en agrupamiento) si no pertenece fuertemente a ningún grupo.

Una ventaja de este tipo de técnicas respecto a las previamene vistas, es que se puede escoger un algoritmo de agrupamiento con una baja complejidad en tiempo y memoria para así hacer la detección de anomalías bastante eficiente. No obstante, algunos algoritmos de agrupamiento requieren el número de grupos como parámetro, lo que puede ocasionar que un objeto sea atípico o no dependiendo de dicho número. Repetir el análisis para diferentes números de clases puede ayudar a tratar con este inconveniente.

Capítulo 3

Aprendizaje Automático

El aprendizaje automático (o aprendizaje máquina) puede definirse de manera general como un conjunto de métodos computacionales que utilizan experiencia para mejorar su rendimiento o hacer predicciones, donde experiencia hace referencia a la información disponible para aprendizaje. La información puede obtenerse de conjuntos de datos cuya calidad y tamaño es crucial para el éxito de los algoritmos. Esto hace que naturalmente el aprendizaje automático esté relacionado con la estadística y el análisis de datos [Mohri et al., 2012]; las técnicas de aprendizaje automático son métodos basados en datos que combinan conceptos fundamentales de las ciencias de la computación con ideas de estadística, probabilidad y optimización. Los algoritmos de aprendizaje tienen una variedad de aplicaciones en las que podemos incluir la clasificación de textos, procesamiento de lenguaje natural, reconocimiento de voz, reconocimiento de imágenes, biología computacional y diagnósticos médicos. El aprendizaje automático se puede dividir en distintos tipos de acuerdo a los datos que se tengan disponibles para entrenamiento, al método que se aplique en ellos y a los datos que se utilizarán para la evaluación. En adelante nos enfocaremos en dos tipos: aprendizaje supervisado y no supervisado [Murphy, 2012].

Aprendizaje supervisado: Los algoritmos tienen como objetivo crear un modelo que mapee un conjunto de entradas $\{\mathbf{x}\}$ a un conjunto de salidas $\{y\}$, dado un conjunto etiquetado de parejas de entrada y salida, $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$. D es el conjunto de entrenamiento y n es el número de objetos de entrenamiento. Los casos más comunes de aprendizaje supervisado son la clasificación y la regresión. Si las salidas y_i son categóricas o discretas, el problema es de clasificación, mientras que si las y_i son continuas entonces el problema es de regresión.

Aprendizaje no supervisado: En el caso no supervisado, los algoritmos únicamente trabajan con un conjunto de entrada $D = \{\mathbf{x}_i\}_{i=1}^{n}$ y el objetivo es encontrar patrones de interés. El agrupamiento y la reducción de la dimensionalidad son ejemplos de problemas de aprendizaje no supervisado.

En las secciones siguientes se explican algunos de los algoritmos más conocidos para clasificación y agrupamiento, y que son importantes para nuestro trabajo.

3.1. Análisis de Grupos

Como se mencionó previamente, en el aprendizaje automático no supervisado únicamente contamos con un conjunto de datos de entrada, por lo que el interés no es hacer predicciones. La tarea en este caso es encontrar estructuras de interés en los datos, una de ellas es un agrupamiento o *clustering*, lo que consiste en particionar un conjunto de datos $D = \{\mathbf{x}_i\}_{i=1}^{N}$ en un determinado número k de grupos o *clusters*, $C = \{C_1, C_2, \ldots C_k\}$. La partición debe ser tal que los objetos en un grupo sean similares entre ellos y distintos de los objetos de otros grupos [Tan et al., 2013].

Existen varias nociones de lo que es un grupo, lo que permite diferenciar a los algoritmos de agrupamiento que existen. Algunos modelos de agrupamiento son:

Basados en prototipos o representantes: Para cada grupo existe un punto representativo (prototipo) que los encapsula. Este objeto es a menudo la media de los objetos del grupo. El método k-means (y sus variaciones) es el más utilizado de este tipo.

Basados en densidad: Aquí un grupo consiste en una región densa de objetos rodeada por una región de baja densidad. El algoritmo basado en densidad más popular es DBSCAN (agrupamiento espacial basado en densidad de aplicaciones con ruido).

Basados en modelos: Los grupos se modelan mediante distribuciones estadísticas, de manera que los objetos en un grupo tengan alta probabilidad de pertenecer a la misma distribución. El método más empleado es el modelo de mezcla Gaussiana, utilizando el algoritmo de esperanza-maximización.

Jerárquicos: Estas técnicas están basadas en la idea de que los objetos están más relacionados a objetos cercanos que aquellos que están más alejados; conectan objetos para formar grupos de acuerdo a su distancia. El agrupamiento jerárquico puede ser aglomerativo, si se inicia con cada objeto como un cluster individual y en cada paso se van mezclando, o divisivo, si se inicia con un único cluster que se divide en cada paso.

En este documento sólo nos centraremos en dos algoritmos basados en prototipos: k-means y k-medoids.

3.1.1. K-means

La técnica basada en prototipos, K-means, es una de las técnicas de agrupamiento más simples y utilizadas. El algoritmo básico de K-means requiere, para su funcionamiento, especificar el número deseado de grupos k. K-means inicializa los centros de los grupos generando aleatoriamente k puntos en el espacio. Cada iteración consiste en dos pasos [Zaki et al., 2014]:

1. Asignación de los objetos a un cluster.

2. Actualización de los centros.

En el primer paso, cada objeto $\mathbf{x}_j \in D$ es asignado al centro más cercano, lo que induce un conjunto de grupos $C = \{C_1, C_2, \ldots C_k\}$, donde los puntos de cada grupo están más cercanos a su centro que al de cualquier otro. En el segundo paso, los centros se actualizan al valor de la media de los puntos de cada grupo:

$$\mu_i = \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j. \tag{3.1}$$

El objetivo del agrupamiento puede expresarse como una función objetivo que dependa de la proximidad de un punto a otro, o de un punto a un centro. En K-means, para datos en el espacio Euclideano, se considera la distancia cuadrada como función de proximidad; esto es, en el primer paso de cada iteración, cada punto x_j es asignado a un cluster C_j*, donde

$$j* = arg \min_{i=1}^{k} \{||\mathbf{x}_j - \mu_i||\}. \tag{3.2}$$

Como función objetivo, la cual evalua la calidad de los grupos, se utiliza la *suma de los errores cuadrados* (SSE por sus siglas en inglés), definida como

$$SSE(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||\mathbf{x}_j - \mu_i||^2. \tag{3.3}$$

La asignación de grupos y la actualización de los centros se llevan a cabo iterativamente hasta alcanzar un punto fijo a un mínimo local. Se puede asumir que K-means converge si los centros no cambian de una iteración a la siguiente, por lo que podemos detener el algoritmo si $\sum_{i=1}^{k} ||\mu_i^t - \mu_i^{t-1}||^2 < \epsilon$, donde $\epsilon > 0$ es el umbral de convergencia y t denota la iteración en turno. Con todo lo anterior, podemos ahora presentar el pseudocódigo de K-means en el algoritmo 3.1.1 [Zaki et al., 2014].

Dado que el algoritmo comienza con centros aleatorios, K-means se ejecuta varias veces, y posteriormente se elige el agrupamiento que tenga el valor de la SSE más bajo. En cuanto a la complejidad computacional, podemos notar que el paso de asignación de grupos ocupa un tiempo $O(nkd)$, puesto que para cada uno de los n puntos se calculan sus distancias a cada cluster k, lo que toma d operaciones en d dimensiones. El paso de actualización de los centros ocupa un tiempo $O(nd)$, ya que se tiene que sumar un total de n puntos de dimensión d. Asumiento que se realizan t iteraciones, la complejidad total de K-means queda como $O(tnkd)$. En términos de almacenamiento, únicamente se necesitan guardar los puntos y los centros, de manera que el espacio requerido es $O(d(n + k))$.

Un ejemplo de K-means aplicado al conjunto de datos Iris se muestra en la figura 3.1. En esta figura se comparan las clases originales del conjunto de datos con los resultados obtenidos por k-means para 2 y 3 grupos.

Algoritmo 3.1.1: K-means

Input: D, k y ϵ

$t = 0$

Inicializa aleatoriamente k centros: $\mu_1^t, \mu_2^t, \ldots, \mu_k^t \in \mathbb{R}^d$

repeat

  $t \leftarrow t + 1$

  `// Asignación de clusters`

  foreach $\mathbf{x}_j \in D$ **do**

    $j* \leftarrow arg\,\text{mín}_i\{||\mathbf{x}_j - \mu_i^t||^2\}$ `// Asigna` $\mathbf{x}_j$ `al centro más cercano`

    $C_{j*} \leftarrow C_{j*} \cup \{\mathbf{x}_j\}$

  `// Actualización de los centros`

  foreach $i = 1$ *to* k **do**

    $\mu_i^t \leftarrow \frac{1}{|C_i|} \sum_{\mathbf{x}_j \in C_i} \mathbf{x}_j$

until $\sum_{i=1}^{k} ||\mu_i^t - \mu_i^{t-1}|| < \epsilon$

K-means es un método sencillo de entender y de implementar, que puede aplicarse sobre una amplia variedad de tipos de datos [Tan et al., 2013]. Además suele ser bastante eficiente a pesar de que se requiera a menudo de varias ejecuciones. Sin embargo, no puede lidiar con grupos de tamaños o densidades distintas, además de ser susceptible a valores atípicos, por lo que suele ser necesario emplear una ténica de detección de anomalías sobre los datos antes de aplicar K-means. Por último, cabe mencionar que está restringido a datos para los que la noción de centro existe. El método que se presenta a continuación no presenta esta restricción.

3.1.2. K-medoids

Como se mencionó previamente, K-means emplea como medida de disimilitud al cuadrado de la distancia euclideana. Esto requiere que todas las variables del conjunto de datos sean atributos cuantitativos. Estas restricciones pueden retirarse a cambio de un mayor costo computacional [Hastie et al., 2009]. La única parte de K-means que utiliza el cuadrado de la distancia Euclideana es el paso de asignación de grupos, mientras que el representante de cada grupo se toma como la media del grupo. Este procedimiento puede generalizarse con el uso de una medida de disimilitud arbitraria, $d(\mathbf{x}_i, \mathbf{x}_j)$, y restringiendo a los centros de cada grupo a tomar el valor de uno de los objetos pertenecientes a ellos. El algoritmo general se describe en 3.1.2 [Hastie et al., 2009].

El costo del paso de asignación permanece igual que en el algoritmo de K-means, si embargo, el paso de actualización de los centros ahora es de $O(k(n-k)^2)$ en cada iteración del algoritmo. K-medoids es más robusto al ruido que K-means, sin embargo, también es más costoso. La implementación más conocida de K-medoids es PAM (Partición alrededor de medoids), aunque para conjuntos de datos grandes es mejor utilizar *clara* (agrupación de grandes aplicaciones) [Kaufman and Rousseeuw, 2009].

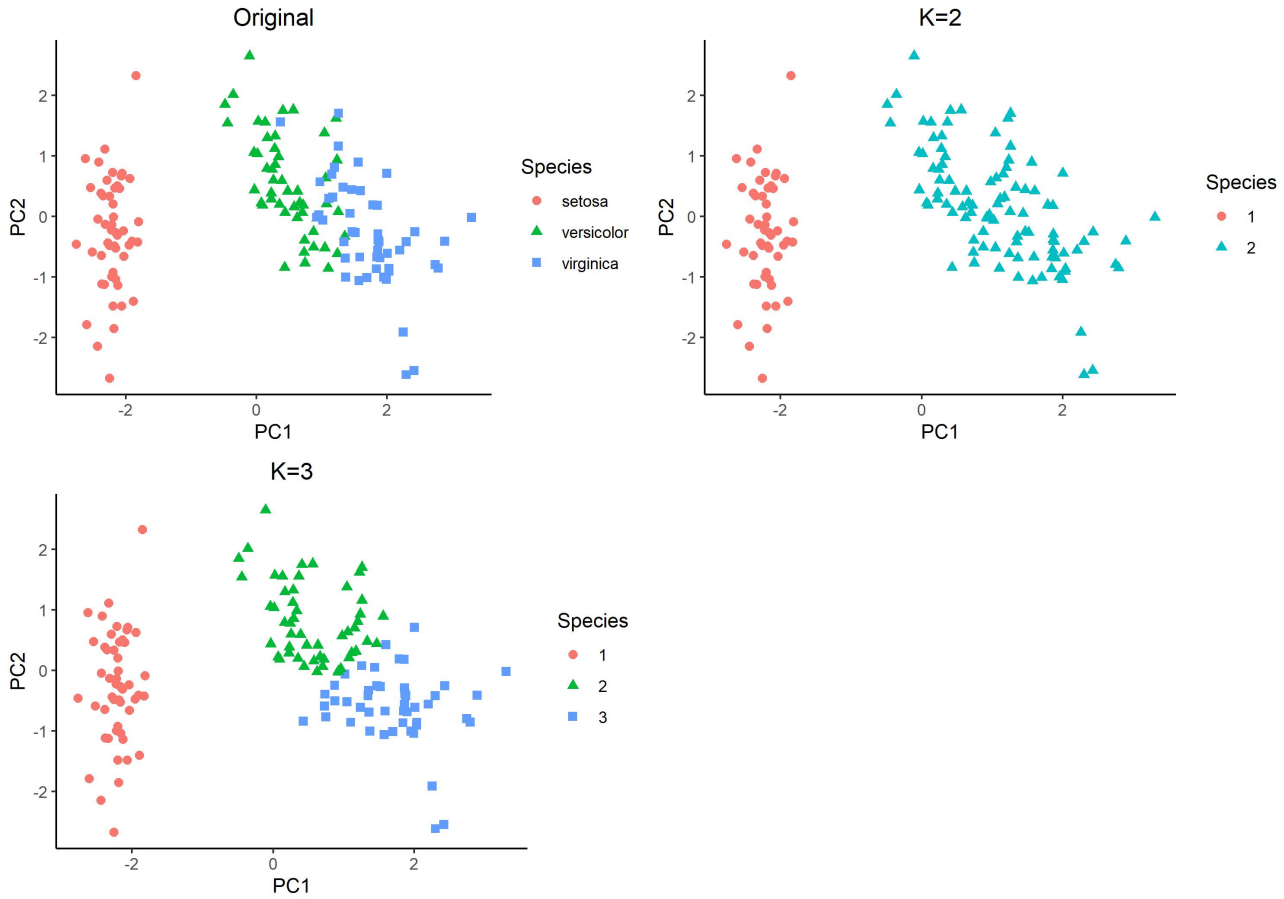

Figura 3.1: Gráficas que muestran las clases originales y la aplicación de k-means sobre el conjunto de datos Iris para $k = 2$ y $k = 3$. PC1 y PC2 son las dos primeras componentes principales.

3.2. Clasificación

La clasificación es una tarea de aprendizaje automático supervisado cuyo objetivo es construir un modelo o una función objetivo que mapee un conjunto de entradas $\{\mathbf{x}_i\}$ a un conjunto de salidas $\{y_i\}$, a partir de un conjunto dado de pares $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ conocido como *conjunto de entrenamiento*. Aquí, $\mathbf{x}_i$ es un vector de atributos o características de dimensión d, mientras que la variable de respuesta y_i es una variable categórica o nominal de algún conjunto finito, $y_i \in \{c_1, \ldots, c_k\}$ [Murphy, 2012]. Existen muchas técnicas de clasificación, a continuación abordaremos algunas de ellas.

3.2.1. Árboles de decisión

Un clasificador de árbol de decisión es un modelo de árbol recursivo basado en particiones que predice la clase y_i para cada punto $\mathbf{x}_i$. Sea R la región que cubre el conjunto de datos D. A rasgos generales, un ábol de decisión utiliza un hiperplano paralelo a

Algoritmo 3.1.2: K-medoids

1 Para un agrupamiento dado, C, encontrar el elemento en cada grupo que minimice la distancia total a otros puntos en el grupo,

$$i*_n \leftarrow arg\,\min_i \sum_{x_j \in C_n} d(\mathbf{x}_i, \mathbf{x}_j).$$

Luego $\mathbf{m}_n \leftarrow \mathbf{x}_{i*_n}$, $n = 1, \ldots, k$ son los actuales centros de los grupos.

2 Dado el actual conjunto de centros $\{\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_k\}$, minizar el error total asignando cada observación al centro más cercano,

$$\begin{aligned} j* &\leftarrow arg\,\min_i d(\mathbf{x}_j, \mathbf{m}_i) \\ C_{j*} &\leftarrow C_{j*} \cup \{\mathbf{x}_j\} \end{aligned}$$

3 Iterar los pasos 1 y 2 hasta que las asignaciones no cambien.

un eje para separar el espacio de datos R en dos, digamos R_1 y R_2, lo que también implica una partición de los puntos de entrada en D_1 y D_2, respectivamente. Cada una de las regiones resultantes es separada recursivamente mediante planos paralelos a los ejes hasta que los puntos dentro de una partición pertenezcan, en su mayoría, a la misma clase. Al final, la jerarquía de las decisiones de separación constituye el modelo del árbol de decisión, con las hojas etiquetadas con la clase mayoritaria entre los puntos en esas regiones. Para clasificar un punto de prueba, se evalúa recursivamente la partición del espacio que le corresponde hasta que se alcance una hoja en el árbol de decisión; la clase a la que pertenece es la etiqueta de la hoja.

Un árbol de decisión se compone de nodos internos que representan las decisiones correspondientes a los hiperplanos o puntos de separación, y de hojas que representan regiones del espacio de datos que están etiquetadas por la clase mayoritaria. Una región está caracterizada por el subconjunto de puntos que pertenecen a ella. Establezcamos algunas definiciones.

Definición 3.2.1. Un *hiperplano* $h(\mathbf{x})$ es el conjunto de todos los puntos $\mathbf{x}$ que satisfacen

$$h(\mathbf{x}) = \mathbf{w}\mathbf{x} + b = 0, \tag{3.4}$$

donde $\mathbf{w} \in \mathbb{R}^d$ es un vector de peso normal al hiperplano y b es el desplazamiento desde el origen del hiperplano.

Dado que un árbol de decisión únicamente considera hiperplanos paralelos a los ejes, el vector de peso se restringe a ser uno de los vectores base $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_d$ de $\mathbb{R}^d$, donde

$\mathbf{e}_i$ tiene un valor de 1 en su i-ésima entrada y 0 en las demás. Si $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$ y suponiendo $\mathbf{w} = \mathbf{e}_j$, entonces la ecuación 3.4 queda como

$$h(\mathbf{x}) = \mathbf{e}_j^T\mathbf{x} + b = 0 \Longrightarrow h(\mathbf{x}) = x_j + b = 0,$$

por lo que la elección de b determina distintos hiperplanos a lo largo de X_j.

Un hiperplano especifíca una decisión o *punto de separación*, pues separa el espacio de datos R en dos. Todos los puntos $\mathbf{x}$ tales que $h(\mathbf{x}) \leq 0$ se encuentran sobre o de una lado del hiperplano, mientras que los que cumplen $h(\mathbf{x}) > 0$ se encuentran al otro lado.

Definición 3.2.2. La forma genérica del *punto de separación* asociado a un hiperplano para un atributo numérico X_j está dado por

$$X_j \leq v,$$

donde $v = -b$ es algún valor en el dominio del atributo X_j.

Entonces, el punto de separación separa el espacio de datos de entrada R en dos regiones R_1 y R_2, que denotan el conjunto de todos los puntos que satisfacen la decisión y el conjunto de los que no la satisfacen, respectivamente.

Definición 3.2.3. La *partición* inducida por un punto de separación de la forma $X_j \leq v$ es

$$\begin{aligned} D_1 &= \{\mathbf{x} : \mathbf{x} \in D, x_j \leq v\}, \\ D_2 &= \{\mathbf{x} : \mathbf{x} \in D, x_j > v\}, \end{aligned}$$

donde D_1 es el subconjunto de datos que pertenecen a la región R_1 y D_2 es el subconjunto de puntos que pertenen a R_2.

Definición 3.2.4. La *pureza* de una región R_j está definida por

$$purity(D_j) = \max_i\{\frac{n_{ji}}{n_j}\},$$

donde $n_j = |D_j|$ (el número de puntos en la región R_j) y n_{ji} es el número de puntos de D_j de la clase c_i.

Podemos ahora presentar el pseudocódigo (algoritmo 3.2.1) para la construcción de un modelo de árbol de decisión [Zaki et al., 2014].

El algoritmo toma como entradas el conjunto de datos en entrenamiento D y dos parámetros η y π; η es el tamaño de las hojas y pi el umbral de pureza. La condición de paro que se impone con el parámetro η previene el sobreajuste del modelo a los datos de entrenamiento, evitando que se modelen subconjuntos muy pequeños. Si una partición es pura (es decir, sobrepasa el umbral de pureza π) entonces no tiene sentido dividirla, por lo que el proceso de partición se termina.

Puede notarse que los árboles de decisión también manejan atributos categóricos. Si X_j es una variable categórica, los puntos de separación son las $X_j \in V$, donde $V \subset dom(X_j)$. A continuación veremos algunas medidas utilizadas para elegir los puntos de separación para atributos numéricos.

Algoritmo 3.2.1: Árbol de decisión

Function `DecisionTree`(D, η, π)**:**
Input: D, η y π
1 $n \leftarrow |D|$ `// tamaño de la partición`
2 $n_i \leftarrow |\{\mathbf{x}_j : \mathbf{x}_j \in D, y_j = c_i\}|$ `// tamaño de la clase` c_i
3 $purity(D) \leftarrow \max_i \frac{n_i}{n}$
4 **if** $n \leq \eta$ *o* $purity(D) \geq \pi$ **then** `// condición de paro`
5 $c* \leftarrow arg\max_i\{\frac{n_i}{n}\}$ `// clase mayoritaria`
6 Crear una hoja y etiquetarla como clase $c*$
7 **return**
8 $(sp*, score*) \leftarrow (\emptyset, 0)$ `// se inicializa el mejor punto de separación`
9 **foreach** *atributo* X_j **do**
if X_j *es numérico* **then**
10 $(v, score) \leftarrow$ `Evaluate-Numerical-Attribute`(D, X_j)
if $score > score*$ **then** $(sp*, score*) \leftarrow (X_j \leq v, score)$
else if X_j *es categórico* **then**
11 $(V, score) \leftarrow$ `Evaluate-Categorical-Attribute`(D, X_j)
if $score > score*$ **then** $(sp*, score*) \leftarrow (X_j \in V, score)$
`//` `Particionar` D `en` D_1 `y` D_2 usando $sp*$ y hacer llamadas recursivas
12 $D_1 \leftarrow \{\mathbf{x} \in D : \mathbf{x}$ satisface $sp*\}$
13 $D_2 \leftarrow \{\mathbf{x} \in D : \mathbf{x}$ no satisface $sp*\}$
14 Se crea el nodo interno $sp*$ con dos nodos hijos, D_1 y D_2
15 `DecisionTree`(D_1, η, π); `DecisionTree`(D_2, η, π)

Medidas para evaluar puntos de separación

Dado un punto de separación $X_j \leq v$, necesitamos algún criterio para evaluarlo. Lo que se busca es escoger aquel punto que otorge la mejor separación entre las diferentes clases. Las medidas más conocidas son la *ganancia* y el *coeficiente de Gini*. Para definir la ganancia, primero necesitamos saber lo que es *entropía*.

Definición 3.2.5. La *entropía* de un conjunto de datos etiquetados D está dada por

$$H(D) = -\sum_{i=1}^{k} P(c_i|D) \log_2 P(c_i|D),$$

donde $P(c_i|D)$ es la probabilidad de la clase c_i en D, y k es el número de clases.

En general, la entropía mide la cantidad de desorden o incertidumbre en un sistema. Si una región es pura (todos los puntos dentro de ella son de una misma clase) entonces tendrá una entropía cero. Por otra parte, si hay una mezcla de todas las clases en

la región con probabilidades $P(c_i, D) = 1/k$, entonces la entropía llega a su máximo, $\log_2 k$. Cuando tenemos una partición D_1 y D_2 mediante un punto de separación, puede establecerse una *entropía de separación* como la suma ponderada de las particiones, más específicamente

$$H(D_1, D_2) = \frac{n_1}{n} H(D_1) + \frac{n_2}{n} H(D_2),$$

donde $n = |D|$, $n_1 = |D_1|$ y $n_2 = |D_2|$.

Para observar si la separación resulta en una entropía reducida promedio, se define la *ganancia de información* como

$$Gain(D, D_1, D_2) = H(D) - H(D_1, D_2). \tag{3.5}$$

A medida que la ganancia aumenta, la entropía disminuye, lo que significa un mejor punto de separación. Por lo tanto, la ganancia de información puede ayudar a elegir la mejor separación.

Definición 3.2.6. El *coeficiente de Gini* está definido como

$$G(D) = 1 - \sum_{i=1}^{k} P(c_i|D)^2.$$

Podemos observar que si la partición es pura, entonces la propabilidad de la clase preponderante es de 1 y la del resto de clases 0, por lo que el coeficiente de Gini es igual a 0. En el caso opuesto, si todas las clases están presentes con igual probabilidad $(P(c_i, D) = 1/k)$, entonces $G(D) = \frac{k-1}{k}$. Es así que valores altos del coneficiente de Gini indican un mayor desorden, y viceversa. El coeficiente ponderado de Gini de una partición generada por un punto de separación puede definirse de la siguiente manera.

$$G(D_1, D_2) = \frac{n_1}{n} G(D_1) + \frac{n_2}{n} G(D_2).$$

La entropía y el coeficiente de Gini pueden utilizarse como medidas de evaluación para puntos de separación, sin embargo, dependen de la función de probabilidad para D $(P(c_i|D))$ y para las particiones D_1 y D_2 $(P(c_i|D_1)$ y $P(c_i|D_2))$. Este último par de funciones deben calcularse para todos los posibles puntos de separación, lo que es computacionalmente costoso. Sin embargo, los puntos de separación de la forma $X \leq v$ producen las mismas particiones para valores $v \in [x_a, x_b)$, donde x_a y x_b son dos valores sucesivos distintos de X en D. Bajo esta idea, las funciones de probabilidad $P(c_i|D_1)$ y $P(c_i|D_2)$ pueden estimar como sigue [Zaki et al., 2014].

$$\hat{P}(c_i|D_1) = \frac{N_{vi}}{\sum_{j=1}^{k} N_{vj}}, \quad \hat{P}(c_i|D_2) = \frac{n_i - N_{vi}}{\sum_{j=1}^{k} (n_j - N_{vj})},$$

donde N_{iv} es el número de puntos $x_j \leq v$ de la clase c_i (x_j es el valor del objeto $\mathbf{x}_j$ en el atributo X_j), $n = |D|$ y n_i el número de puntos en D de la clase c_i. De esta manera, se puede ahora presentar el pseudocódigo (algoritmo 3.2.2) para la evaluación de puntos de separación con variables numéricas.

Algoritmo 3.2.2: Algoritmo de evaluación de atributos numéricos

Function `Evaluate-Categorical-Attribute(`D, X`):`

1 Ordenar D sobre el atributo X de manera que $x_j \leq x_{j+1}, \forall j = 1, \ldots, n-1$
2 $M \leftarrow \emptyset$ `// conjunto de puntos intermedios`
3 **for** $i = 1, \ldots, k$ **do** $n_i \leftarrow 0$
4 **for** $j = 1, \ldots, n-1$ **do**
5 **if** $y_j = c_i$ **then** $n_i \leftarrow n_i + 1$
6 **if** $x_{j+1} \neq x_j$ **then**
7 $v \leftarrow \frac{x_{j+1}+x_j}{2}$; $M \leftarrow M \cup v$ `// puntos intermedios`
8 **for** $i = 1, \ldots, k$ **do**
9 $N_{vi} \leftarrow n_i$ `// número de puntos tales que` $x_j \leq v$ `y` $y_j = c_i$
10 **if** $y_n = c_i$ **then** $n_i \leftarrow n_i + 1$
`// evalúa los puntos de separación de la forma` $X \leq v$
11 $v* \leftarrow \emptyset$; $score* \leftarrow 0$ `// inicializa el mejor punto de separación`
12 **forall** $v \in M$ **do**
13 **for** $i = 1, \ldots, k$ **do**
14 $\hat{P}(c_i|D_1) \leftarrow \frac{N_{vi}}{\sum_{j=1}^{k} N_{vj}}$
15 $\hat{P}(c_i|D_2) \leftarrow \frac{n_i - N_{vi}}{\sum_{j=1}^{k} n_j - N_{vj}}$
16 $score(X \leq v) \leftarrow$ `Gain(`D, D_1, D_2`)`
17 **if** $score(X \leq v) > score*$ **then**
18 $v* \leftarrow v$; $score* \leftarrow score(X \leq v)$
19 **return** $(v*, score*)$

En cuanto a la complejidad de este algoritmo, la línea 1 toma un tiempo $O(nlogn)$ y la línea 4 $O(nk)$. El costo del ciclo de la línea 12 también tiene un costo $O(nk)$, ya que el número de puntos intermedios puede ser a lo sumo n. Entonces la complejidad total del algoritmo 3.2.2 es $O(nlogn + nk)$, pero como k suele ser una constante con valor bajo, el costo todal puede quedar como $O(nlogn)$.

El algoritmo 3.2.1 evalúa los puntos intermedios para cada uno de los d atributos. El costo de este algoritmo depende de la profundidad del árbol, la cual en el peor de los casos será de n. Así la complejidad total es $O(dn^2logn)$.

En la figura 3.2 se grafican los puntos del conjunto de datos Iris respecto a los atributos *sepal.lenght* y *sepal.width*. Se ilustran las separaciones hechas por los hiperplanos paralelos a los ejes. El árbol de decisión inducido por este particionamiento se muestra

en la figura 3.3. El primer hiperplano es $h_1(\mathbf{x}) = x_1 - 5{,}45 = 0$, el cual corresponde a la decisión *sepal.lenght* $\leq 5{,}4$. Posteriormente, las dos regiones generadas son ahora divididas por los hiperplanos $h_2(\mathbf{x}) = x_2 - 2{,}8$ y $h_3(\mathbf{x}) = x_2 - 3{,}4$ correspondientes a las decisiones *sepal.width* $\leq 2{,}8$ y *sepal.width* $\leq 3{,}4$, respectivamente.

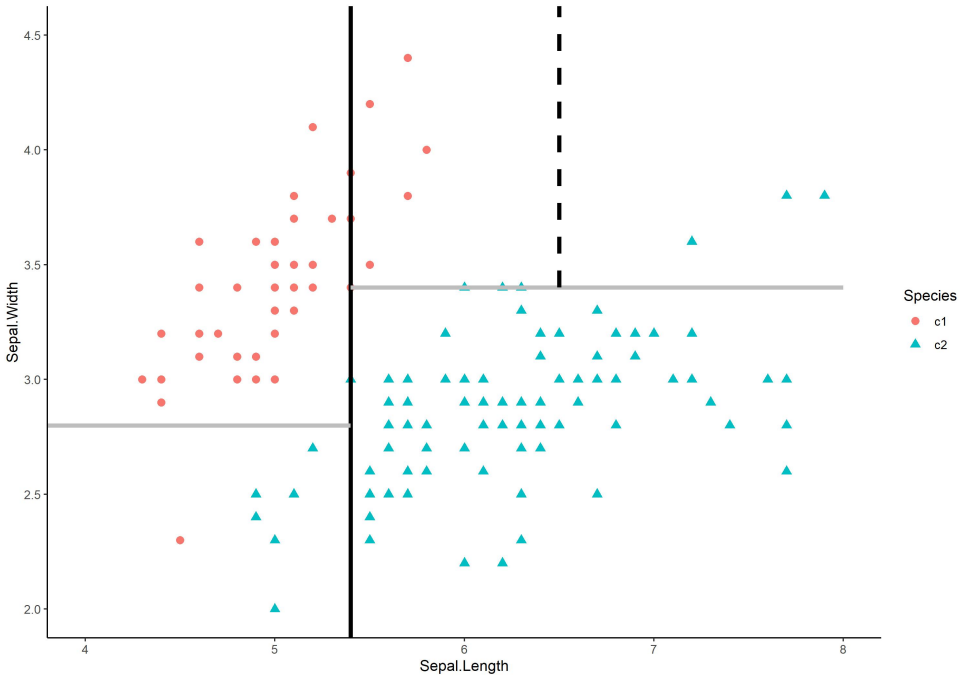

Figura 3.2: Separaciones por hiperplanos.

Los árboles de decisión son fáciles de explicar y pueden presentarse gráficamente, lo que los hace fáciles de interpretar. Sin embargo, generalmente suelen ser menos eficientes en cuanto a precisión a diferencia de otros métodos de clasificación [James et al., 2013].

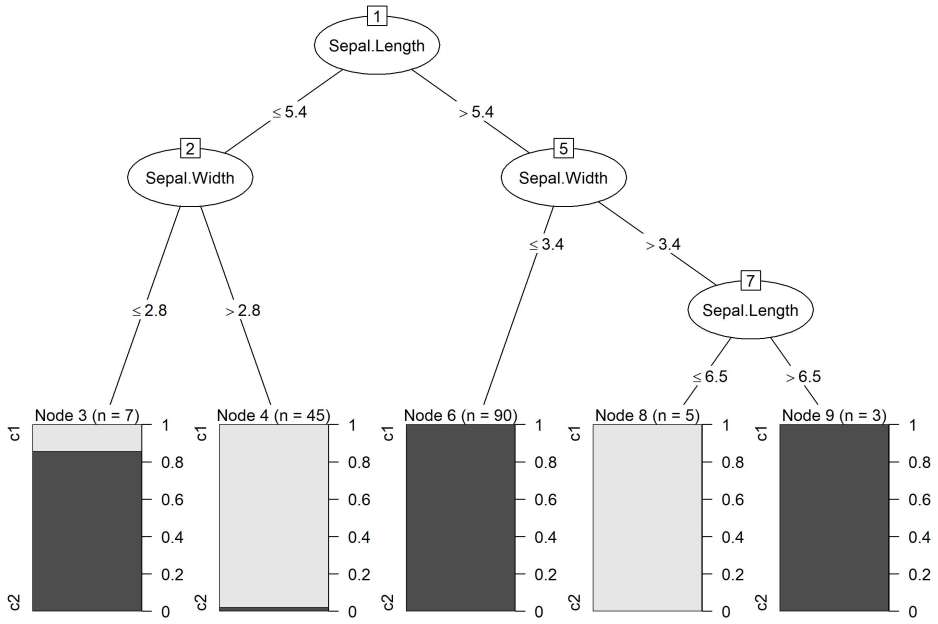

Figura 3.3: Árbol de decisión

3.2.2. Máquinas de vectores de soporte

Sea $D = (\mathbf{x}_i, y_i)_{i=1}^n$ un conjunto de datos de clasificación, con n puntos de dimensión d, y supongamos que únicamente existen dos clases, es decir, $y_1, \ldots, y_n \in \{-1, 1\}$. La clasificación mediante *máquinas de vectores de soporte* o *máquinas de soporte vectorial* (SVM por su siglas en inglés) tiene como base separar los puntos $\mathbf{x}_i$ mediante un hiperplano de $d-1$ dimensiones. El hiperplano debe elejirse de forma que maximice el margen de separación entre las dos clases. Una ilustración del funcionamiento de una máquina de soporte vectorial se muestra en la imagen 3.4

Para cada punto $\mathbf{x}_i$ se puede calcular su distancia a un hiperplano de separación $h(\mathbf{x})$ mediante

$$\delta_i = \frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{||\mathbf{w}||}.$$

Se define el *margen* del clasificador lineal como la mínima distancia de un punto al hiperplano:

$$\delta* = \min_{\mathbf{x}_i}\{\frac{y_i(\mathbf{w}^T\mathbf{x}_i + b)}{||\mathbf{w}||}\}.$$

Todos los puntos $\mathbf{x}*$ que cumplan este requisito se conocen como *vectores de soporte* para el hiperplano. Si el hiperplano de separación se reescala de forma que $y*h(\mathbf{x}*) = 1$,

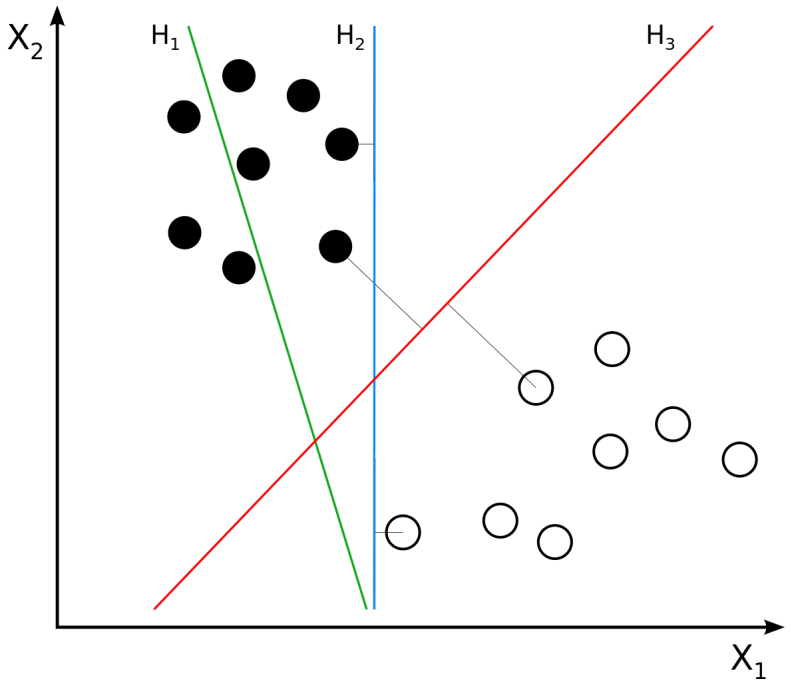

Figura 3.4: Gráfica que muestra la forma en que una máquina de vectores de soporte elige un hiperplano de separación; el plano H_3 separa los puntos con el margen de separación máximo. Tomada de [Commons, 2012].

entonces el margen queda como

$$\delta* = \frac{1}{||\mathbf{w}||}.$$

Para todos los puntos $\mathbf{x}_i$ que no sean vectores de soporte se tiene que $y_i h(\mathbf{x}_i) > 1$, ya que por definición deben encontrarse más lejos del hiperplano que un vector de soporte. Entonces, para todos los puntos del conjunto de datos D, se obienen las siguientes desigualdades

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall \mathbf{x}_i \in D.$$

Las máquinas de soporte vectorial requieren la solución del siguiente problema de optimización [Zaki et al., 2014]:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{n}(\xi_i)^k \\ \textbf{sujeto a} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i \; \forall \mathbf{x}_i \in D, \end{aligned} \tag{3.6}$$

donde C y k son dos parámetros de penalización para el término de error ξ, $\xi_i \geq 0 \forall \mathbf{x}_i \in D$. Si $\xi_j = 0$ significa que el punto x_j está alejado al menos un distancia $\frac{1}{||\mathbf{w}||}$ del hiperplano. Si $0 < \xi_j < 1$ entonces $\mathbf{x}_j$ se encuentra sobre el hiperplano, por lo que

aún está correctamente clasificado. Sin embargo, si $\xi_j > 1$ entonces el punto está mal clasificado y se encuentra en el lado erróneo del hiperplano. Si $\xi_i = 0\ \forall \mathbf{x}_i$, los puntos son linealmente separables. En este caso se omite el uso de los parámetros C y k, y de los términos ξ en el problema de optimización 3.6.

Kernel SVM

El enfoque lineal para las SVM puede generalizarse para conjuntos de datos no lineales [Boser et al., 1992]. La idea es mapear el conjunto original de puntos $\mathbf{x}_i$ de d dimensiones a puntos $\phi(x_i)$ de un espacio de características mediante alguna transformación no lineal ϕ. Esto se hace reemplazando el producto interior por una función núcleo o kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j).$$

De esta manera, el problema de optimización para el caso no lineal queda como:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{n}(\xi_i)^k \\ \textbf{sujeto a} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i\ \forall \mathbf{x}_i \in D. \end{aligned} \tag{3.7}$$

Algunos de los kernels más utilizados [Hsu et al., 2003] son:

Polinomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j)^p,\ \gamma > 0.$

Función de base radial: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2).$

Tangente hiperbólica: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \mathbf{x}_i^T \mathbf{x}_j + r).$

Aquí γ, r y p son parámetros del kernel.

SVM para más de dos clases

Hasta ahora se ha hablado únicamente de máquinas de soporte vectorial para clasificación binaria. Existen varias maneras de extender las SVM para k clases, las más conocidas [James et al., 2013] son

Uno contra uno Este enfoque construye $\binom{k}{2}$ SVM, cada una comparando un par de clases. Para clasificar un objeto de prueba, se emplean todos los $\binom{k}{2}$ construidos y se cuenta el número de veces en las que el objeto es asignado a cada una de las clases; la clase con mayor frecuencia es la que se asigna al objeto.

Uno contra todos Aquí se entrenan k SVM, cada vez comparando una de las k clases con las $k-1$ restantes. Los objetos de la clase k se consideran positivas, y el resto negativas

3.2.3. Redes Neuronales Artificiales

La inspiración de las redes neuronales artificiales fue el reconocimiento de sistemas de aprendizaje complejos en los cerebros de los seres vivos, compuesto de un conjunto interconectado de neuronas [Larose and Larose, 2014]. Un nodo o neurona se puede modelar mediante las señales de entrada x_i, un vector de pesos $\mathbf{w}$ y una función de activación σ.

Entonces el valor de salida de una neurona viene dado por

$$y = \sigma(\sum_i w_i x_i).$$

Aunque una neurona pueda parecer simple en cuanto a estructura, las redes de neuronas interconectadas pueden realizar tareas complejas como clasificación y reconocimiento de patrones. En esta sección describiremos las bases de la red neuronal más ampliamente utilizada, denominada red de propagación hacia atrás o red neuronal prealimentada. Este tipo de red se puede modelar mediante tres capas (como en la figura 3.5): la de entrada, la oculta y la de salida. Para una clasificación de k clases hay k nodos en la capa de salida, con el j-ésimo nodo modelando la probabilidad de la clase j. Cada nodo de entrada representa una de las d características de un objeto de entrada. Si se tienen m nodos en la capa oculta, entonces las características derivadas z_i en cada uno de ellos se crean a partir de combinaciones lineales de las entradas [Hastie et al., 2009], es decir

$$z_i = \sigma(\alpha_{i0} + \sum_{j=1}^{d} \alpha_{ij} x_j),\ i = 1, \ldots, m.$$

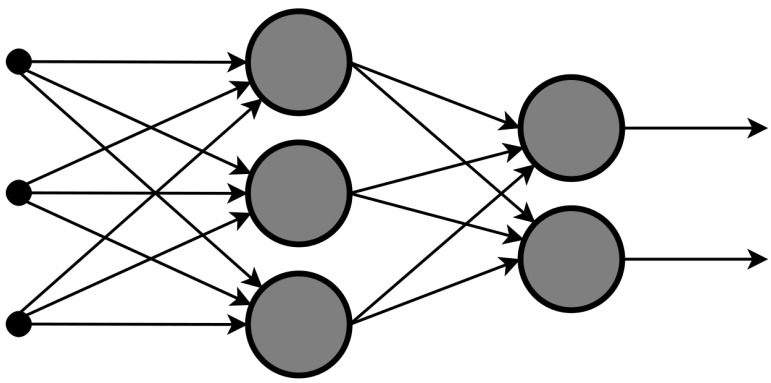

Figura 3.5: Ejemplo de una red neuronal con una capa de entrada, una oculta y una de salida. Tomada de [Commons, 2015].

Mientras que en los nodos de la capa de salida, los objetivos y_i se modelan como una función de la combinación lineal de las z_j, esto es

$$y_i = f(\beta_{i0} + \sum_{j=1}^{m} \beta_{ij} z_j),\ i = 1, \ldots, k.$$

Como función se activación se suele utilizar la función sigmoide (ecuación 3.8) ya que esta combina comportamientos casi lineales, curvilineos y casi constantes, dependiendo del valor de v [Larose and Larose, 2014].

$$\sigma(v) = \frac{1}{1 + \exp(-v)}. \tag{3.8}$$

La función de salida f permite una transformación final del vector de salida $t_i = \beta_{i0} + \sum_{j=1}^{m} \beta_{ij} z_j$, esta suele ser la función *softmax*:

$$f(t_i) = \frac{\exp(t_i)}{\sum_{j=1}^{k} \exp(t_j)}.$$

Para el entrenamiento de una red neuronal, cada objeto del conjunto de datos de entrenamiento es procesado a través de la red y se producen k salidas que representan la probabilidad de que el objeto pertenezca a cada una de las k clases. Estas salidas se comparan con el valor de la clase verdadera a la que pertenece el objeto para obtener el error. La función E para la medición del error puede ser la suma de errores cuadráticos:

$$E(x) = SSE = \sum_{j=1}^{k} (y_i' - y_i)^2,$$

o bien, la entropía cruzada:

$$E(x) = \sum_{j=1}^{k} y_i' \log(y_i),$$

donde y_i' suele ser 1 si la clase del objeto es la correcta y 0 en los demás casos. La propagación hacia atrás consiste en ajustar los pesos α y β tal que minimicen el error E. El problema de optimización se resuelve comúnmente con el método del descenso del gradiente, aunque puede usarse otro método de optimización [Ripley, 1996].

Algunas consideraciones a tomar en cuenta de una red neuronal de propagación hacia atrás, son:

1. Dadas las características de la función sigmoide, es recomendable inicializar los pesos con valores aleatorios cercanos a cero. De esta manera el modelo empieza con un comportamiento lineal, y deja de serlo cuando los pesos comienzan a aumentar su valor. Empezar con pesos grandes puede producir soluciones poco óptimas.

2. Si hay demasiados pesos la red neuronal tenderá a sobreajustar los datos al mínimo global de E, por lo que se suele establecer algún criterio de paro.

3. Si hay pocos nodos en la capa oculta, el modelo no podrá manejar muy bien los datos no lineales; con demasiados nodos, los pesos extra pueden ser reducidos a cero. Se recomienda emplear un número de entre 5 y 100 nodos, con cantidades más grandes cuando se tengan también muchos atributos y objetos de entrenamiento.

4. Aunque no es necesario, estandarizar los datos antes de entrenar la red neuronal puede producir mejores resultados.

3.2.4. Bagging y Boosting

Los árboles de decisión discutidos anteriormente tienen problemas de alta varianza [James et al., 2013]. Esto quiere decir que si el conjunto de datos de entrenamiento es dividido en dos partes de manera aleatoria, los resultados obtenidos al entrenar un árbol con ambas partes pueden ser distintos. Por otra lado, los procedimentos con una baja varianza produce resultados similares si se aplica repetidamente a conjuntos de datos distintos. *Bagging* es un procedimiento para reducir la varianza de un método de aprendizaje, que generalmente es muy útil aplicándolo en de árboles de decisión. Una manera natural de reducir la varianza (y así mejorar la exactitud) de un método de aprendizaje es tomar varios conjuntos de entrenamiento, construir un modelo predictivo para cada uno de ellos y promediar las predicciones resultantes. Sin embargo, esto es poco práctico ya que usualmente no disponemos de muchos conjuntos de datos para entrenamiento. Lo que podemos hacer es tomar múltiples muestras de un solo conjunto de datos, para así tener B conjuntos de entrenamiento diferentes. De esta manera, la predicción resultante para un objeto $\mathbf{x}$ quedaría como

$$f_{bag} = \frac{1}{B}\sum_{i=1}^{B} f^i(\mathbf{x}).$$

Bagging puede hacer mejoras en la exactitud combinando cientos o incluso miles de árboles en un solo procedimiento. El número de árboles B no es un parámetro crítico; aunque se utilice un número muy grande de árboles, no se produce sobreajuste.

Mientras bagging trabaja con árboles sobre muestras del conjunto de datos de manera independiente, en *boosting* se trabaja con árboles secuenciales. Boosting no toma muestras de datos, sino que modela los árboles sobre versiones modificadas del conjunto de datos original. Dado un modelo, se ajusta un árbol de decisión a los residuos de este en lugar de sus resultados. Cada nuevo árbol puede ser tan pequeño como para tener sólo dos nodos terminales; esto queda determinado mediante un parámetro γ. Ajustar árboles pequeños a los residuos permite mejorar la predicción. Se introduce también un parámetro de contracción λ que permite ralentizar el proceso, permitiendo que más y diferentes árboles traten con los residuos. Utilizar un valor grande de λ requiere de un valor grande de B (el número de árboles) para conseguir un buen rendimiento. Sin embargo, a diferencia de bagging, boosting tiende a sobreajustar el modelo si B es muy grande.

3.2.5. Bosques Aleatorios

Los bosques aleatorios funcionan bajo el mismo principio que el bagging [James et al., 2013], es decir, modelando árboles de decisión sobre muestras del conjunto original de

datos de entrenamiento. La diferencia radica en que, cada vez que ocurre una separación en el árbol, se elige una muestra aleatoria de m predictores (atributos) de los p totales. El procedimiento se limita a que el punto de separación sea uno de los m candidatos; en cada separación se genera aleatoriamente un nuevo conjunto de atributos. Típicamente se escoge el valor de m tal que $m \approx \sqrt{p}$. Es así que, cuando se construye un bosque aleatorio, al algoritmo no se le permite encontrar el punto de separación que proporcione la mejor separación entre clases. Aunque esto parezca contraproducente, tiene su razón de ser. Si hay una variable con muy alto poder de predicción, la mayoría de los árboles construídos tendrán a esta variable en su raíz; consecuentemente, todos los árboles generarán resultados altamente correlacionados. Promediar muchas variables con alta correlación no otorga una gran reducción a la varianza. En casos como estos, utilizar bagging no reducirá en gran medida la varianza. Si $m = p$, entonces el bosque aleatorio equivale al bagging, por lo que este último (aplicado en árboles) es un caso especial de bosque aleatorio [Tan et al., 2013]. Valores pequeños para m suelen ser ayuda en los casos que se tengan grandes números de variables correlacionadas.

El pseudocódigo de un bosque aleatorio se describe en el algoritmo 3.2.3. Los parámetros que requiere son el número de árboles B, el tamaño de la muestra N, el número de predictores m y el tamaño mínimo de nodos n_{min} [Hastie et al., 2009].

Algoritmo 3.2.3: Bosque aleatorio para clasificación

1 **Input:** B, m, D, N, n_{min}
2 **for** $i = 1, \ldots, B$ **do**
3 Selecciona una muestra $D*$ de tamaño N del conjunto D
4 Genera un árbol de decisión T_i para $D*$, aplicando recursivamente los siguientes pasos, hasta alcanzar n_{min}
- Selecciona m variables de manera aleatoria del total de variables.
- Escoge el mejor punto de separación de las m variables.
- Divide el nodo en dos.

5 **Output:** El ensamble de árboles $\{T_i\}_{i=1}^{B}$

Los métodos de bagging, boosting y bosques aleatorios otorgan una mejora en las predicciones respecto a los árboles de decisión. Sin embargo, a diferencia de estos, la interpretación de los modelos se hace más complicada. Aunque un ensamble de árboles sea más difícil de interpretar que uno solo, se puede obtener un promedio de la importancia de cada variable predictiva [James et al., 2013]. En cada separación de cada árbol, la mejora en el criterio de separación es la medida de la importancia que se le atribuye a la variable separadora; esta es acumulada sobre todos los árboles del ensamble.

3.2.6. Medidas de evaluación

Una vez que un modelo de clasificación M se ha entrenado, podemos medir su rendimiento sobre un conjunto de datos de prueba para los cuales se conocen las clases originales. De esta forma se pueden comparar múltiples clasificadores. Sea y_i la clase verdadera y $\hat{y}_i = M(\mathbf{x}_i)$ la clase predicha para un objeto $\mathbf{x}_i$ del conjunto de prueba. Entonces podemos definir las siguientes medidas [Zaki et al., 2014].

Definición 3.2.7. La *tasa de error* es la fracción de predicciones incorrectas hechas por el clasificador sobre el conjunto de datos de prueba, determinado por

$$err = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i),$$

donde I toma valor 1 si su argumento es verdadero y 0 en caso contrario. Entre menor sea la tasa de error, mejor es el clasificador.

Definición 3.2.8. La *exactitud* de un clasificador es la fracción de predicciones correctas sobre el conjunto de datos de prueba, determinado por

$$acc = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) = 1 - err,$$

La exactitud proporciona un estimado de la probabilidad de una predicción correcta, por lo que el clasificador será mejor a medida que la exactitud sea mayor.

Las medidas descritas anteriormente son globales y no toman en cuenta las clases que contribuyen al error. Se pueden obtener medidas más informativas con ayuda de una *matriz de confusión* N. Esta matriz está compuesta por una partición $O = \{O_1, \ldots, O_k\}$ de los puntos de prueba de acuerdo a las clases originales, y una partición $R = \{R_1, \ldots, R_k\}$ de los puntos respecto a las clases predichas, esto es

$$N(i,j) = n_{ij} = |R_i \cap O_j|,\ R_i = \{\mathbf{x}_l \in D : y_l = c_i\},\ O_j = \{\mathbf{x}_l \in D : \hat{y}_l = c_j\}.$$

Aquí n_{ij} representa el número de puntos con clase predicha c_i y clase original c_j, por lo que n_{ii} representa el número de objetos que el clasificador asignó a la clase original c_i.

Definición 3.2.9. La *precisión* de un clasificador para la clase $c_i,\ i = 1, \ldots, k$, está dada como la fracción de predicciones correctas sobre todos los puntos predichos para la clase c_i :

$$prec_i = \frac{n_{ii}}{|R_i|}.$$

Definición 3.2.10. El *recuerdo* de un clasificador para la clase $c_i,\ i = 1, \ldots, k$ es la fracción de predicciones correctas sobre todos los puntos de la clase c_i:

$$rec_i = \frac{n_{ii}}{|O_i|}.$$

Valores mayores para la precisión y el recuerdo representan mejores clasificadores. Por último definiremos la medida F (F-score), que es un balanceo de la precisión y el recuerdo.

Definición 3.2.11. *F-score* está determinado por la media armónica de la precisión y el recuerdo para la clase c_i

$$F_i = \frac{2 \cdot prec_i \cdot rec_i}{prec_i + rec_i} = \frac{2n_{ii}}{|R_i| + |O_i|}.$$

El F-score general para un clasificador es la media de los valores por clase:

$$F = \frac{1}{k} \sum_{i=1}^{k} F_i.$$

Capítulo 4

Nivel Socioeconómico

Las sociedades humanas siempre han desarrollado estructuras sociales que diferencian a grupos particulares de acuerdo a características que son muy valoradas dentro de estas. Dentro del campo de la sociología, Karl Marx (1818-1883) y Max Weber (1864-1920) fueron dos figuras importantes que establecieron bases para estructurar y comprender las clases sociales. Marx categorizó la clase en función de la relación de un grupo con los medios de producción, enfatizando la desigualdad económica. Él fue muy crítico con el sistema capitalista; el capitalismo es un sistema de producción de mercancías en el cual las personas se involucran en un proceso que satisface tanto sus necesidades como las de sus inmediatos. Es así que las clases se constituyen con la relación entre grupos de aquellos que poseen bienes en los medios de producción y aquellos que no [Lynch and Kaplan, 2007].

Weber, por otra parte, tenía otro punto de vista respecto a la clase social. Para él, las diferencias sociales están basadas en tres características: clase, estatus y poder. Clase implica la posesión y el control de recursos, y queda determinada mediante los ingresos; el estatus hace referencia a factores sociales y culturales como estilo de vida, historial familiar y conexiones sociales; por último, Weber reconoció que tener prestigio y ventajas económicas otorgan un mayor poder [Liberatos et al., 1988].

Las teorías de Marx y Weber han servido a muchos sociólogos para desarrollar nuevos modelos de estratificación de clases sociales, muchos de ellos utilizados en numerosos estudios recientes. Algunos de estos estudios muestran la relación que existe entre el nivel socioeconómico con diversos problemas físicos y mentales ([Werner et al., 2007] [McLaren, 2007]), y con el desarrollo y crecimiento infantil ([Bradley and Corwyn, 2002]). La información que proporciona este indicador social acerca de la población, es también muy utilizada en el área de epidemología, como se explica en [Oakes and Rossi, 2003]. Conocer las características sociales y económicas de una determinada región, puede ayudar a las empresas a decidir en dónde expandirse de acuerdo a sus necesidades.

4.1. Definición y medición

El nivel socieconómico al igual que estructura y clase social, son temas centrales para las ciencias sociales; sin embargo, los sociólogos no han llegado a un acuerdo para sus definiciones. En [Liberatos et al., 1988] se concluye que tampoco existe una mejor medición del nivel socioeconómico. Los investigadores generalmente manejan distintos enfoques dependiendo del modelo de estudio y de la información disponible, sin embargo, algunos de los indicadores comúnmente utilizados y los recomendados por algunos estudios como [Galobardes et al., 2006] son educación, ingresos, ocupación y condiciones de vivienda.

La educación es considerada un indicador crucial puesto que otorga información sobre el potencial del ingresos a lo largo de la vida, mientras que la ocupación y los ingresos proporcionan una medida instantánea de la situación social y económica de los individuos [Shavers, 2007]. El ingreso familiar bruto es la medida más utilizada para el indicador de ingresos y es, a menudo, reportado en categorías (ingreso bajo, medio o alto) en lugar de una variable continua. La ocupación, es un indicador tradicional del nivel socioeconómico porque transmite información respecto al poder, ingresos y educación de una persona, asociados con posiciones dentro de la estructura ocupacional. En cuanto a las condiciones de vivienda, es común utilizar características de tenencia de vivienda (si es propia o rentada por un propietario social o privado), infraestructura (materiales de construcción, número de habitaciones), amenidades (como agua caliente, estufa, lavadora y refrigador) y hacinamiento (sobrepoblación de habitantes en la vivienda).

A continuación presentaremos algunos antecendentes referentes al nivel socioeconómico, con un énfasis en los trabajos realizados en México, ya que estos constituyen una fuerte base para nuestro trabajo.

4.2. Antecedentes

En los Estados Unidos, se han propuesto varios estudios de nivel socieconómico, entre los que podemos destacar el de [Berzofsky et al., 2007]. Este estudio tiene el propóstito de mejorar el análisis de victimización y sus correlaciones mediante ingresos, educación y ocupación de la población como principales indicadores. En el Reino Unido, la Estadística Nacional de Clasificación Socioeconómica (NS-SEC por sus siglas en inglés) es la clasificación oficial del país. En ella se realiza una clasificación social anidada mediante el esquema desarrollado por Goldthorpe, el cual está basado en el modelo de estratificación de Weber.

En México, el INEGI es el organismo público encargado de captar y difundir información del país referente a territorio, recursos, población y economía. La estadística del país la obtienen de tres tipos de fuentes: censos (de población y vivienda, económicos y agrícolas), encuestas (en hogares y en establecimientos) y registros administrativos. Esta misma institución ha elaborado estudios socioeconómicos en distintas ocasiones,

la última de ellas en el año 2004 con un proyecto titulado "Regiones socioeconómicas de México"[INEGI, 2004], mediante el cual clasificaron a las regiones del país dentro de siete estratos distintos, donde el estrato 7 representa a las regiones con la mejor situación respecto a las demás, y el estrato 1 a las regiones más desfavorables; todo esto a nivel estatal, municipal y de área geoestadísica básica. Los resultados que se obtuvieron pueden aún consultarse en su sitio web [INEGI, d].

Para este estudio, el INEGI empleó como fuente de información el *censo de población y vivienda 2000*, construyendo indicadores que describieran características referidas a aspectos de la educación, salud, empleo y vivienda. La selección de los indicadores consistió en cuatro etapas; en la primera se revisaron los indicadores que ellos utilizaron en trabajos previos para decidir cuáles tomar en cosideración, además de proponer algunos nuevos; en la segunda etapa, descartaron variables que pudieran expresar el mismo propósito; en la tercera, emplearon análisis factorial y análisis de correlación para desechar variables que pudieran resultar redundantes; por último, mediante análisis discriminante lineal, se determinaron qué factores tenían o no influencia en la conformación de los estratos. Los indicadores resultantes se muestran en el apéndice A.

El método de clasificación que el INEGI usó como base fue el propuesto por [Jarque, 1981]: se busca inicialmente agrupar los elementos cercanos entre sí en un espacio de d dimensiones mediante la distancia euclideana y posteriormente, con el uso de una función de criterio, se reclasifican los elementos de manera que ningún punto se encuentre más cerca del centro de otro grupo distinto de aquel al que pertenece. El algoritmo requiere de la estandarización de los datos y de los centros (las medias) de cada uno de los grupos inicialmente conformados. Si una observación se reasigna, los centros de los estratos involucrados se actualizan, por lo que cuando el algoritmo termina, se produce un conjunto de centros finales $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_7\}$.

Los algoritmos de agrupamiento, como el que empleó INEGI, no ordenan a los grupos resultantes. Sin embargo, dado que todos los indicadores se construyeron en "sentido positivo"(un mayor valor representa una mejor situación), puede utilizarse el conjunto de centros finales para ordenar a los grupos mediante el siguiente procedimiento [INEGI, 2004]:

1. Se calculan todas las distancias posibles ($\binom{7}{2} = 21$) entre cada pareja de centros finales:

$$d(\mathbf{c}_i, \mathbf{c}_j) = \sum_{l=1}^{d} (c_{il}, c_{jl})^2,$$

   donde $\mathbf{c}_i = (c_{i1}, c_{i2}, \ldots, c_{id})$ y d es el número de atributos. $d(\mathbf{c}_i, \mathbf{c}_j)$ representa la distancia del centro del grupo i al centro del grupo j, $i, j = 1, \ldots, 7$.

2. Se identifica la mayor distancia $d(\mathbf{c}_n, \mathbf{c}_m)$. Entonces el grupo n representa al estrato con la mayor ventaja relativa y el grupo m al estrato con mayor desventaja, o viceversa. Una forma de identificarlos es a través de la lectura de las variables. El centro que reporte los valores más altos en la mayoría de las variables implica

que se trata del estrato que está en la mejor situación relativa y por lo tanto se identifica con el número 7 (el más alto), en consecuencia al otro centro se le identifica con el número 1 (el más bajo).

3. Para ordenar los estratos restantes, se calculan las distancias respecto a alguno de los extremos. Si por ejemplo se elige al centro que representa a los que están en mejor situación (estrato 7), entonces las distancias de los centros restantes a este centro se ordenan de menor a mayor. Por lo tanto la menor distancia (la más cercana) identifica al centro que corresponde al estrato 6, la segunda menor distancia identifica al centro que corresponde al estrato 5 y así sucesivamente. Si por el contrario se elige al centro que representa a los que están en la situación menos favorable (estrato 1), entonces la distancia más cercana identifica al centro que corresponde al estrato 2 y así sucesivamente.

La AMAI es otro organismo que, entre sus diversas labores, realiza constantemente estudios socioeconómicos en el país, [AMAI, 2018] es el último de ellos. El modelo que utilizan se genera a partir de los resultados de la Encuesta Nacional de Ingresos y Gastos de los Hogares (enigh), en la que se registra no sólo información del gasto y el ingreso en los hogares, sino también características de infraestructura de la vivienda que ocupan, características de los integrantes del hogar e incluso algunos datos interesantes acerca de sus hábitos de alimentación.

Para construir su modelo de nivel socioeconómico realizaron un análisis descriptivo de asociaciones sobre las variables tomadas de la enigh, buscando identificar variables que presentan correlaciones interesantes que pudieran ser probadas en el modelo predictivo. Después de el análisis, encontraron que las variables que presentaron el mayor poder predictivo son:

- Nivel educativo del jefe del hogar.
- Número de baños completos en la vivienda.
- Número de autos en el hogar.
- Tenencia de conexión a internet en el hogar.
- Número de integrantes en el hogar mayores de 14 años que trabajan.
- Número de dormitorios en la vivienda.

Para hacer la clasificación la AMAI estableció un sistema de puntaje, en el que un puntaje máximo se reparte de manera proporcional entre cada variable utilizada, para que luego en cada variable se pueda asignar cierto puntaje para cada valor que esta pueda tomar. Una vez asignado el puntaje total correspondiente a cada hogar encuestado en la enigh, se utiliza el procedimiento de estratificación univariado de Dalenius-Hodges para obtener los puntos de corte que minimizan la variabilidad intragrupos. También

utilizan siete estratos para clasificar y presentan sus resultados a nivel de entidad federativa y zona matropolitana.

Hemos mencionado que INEGI elaboró su estudio a nivel estado, municipio y de área geoestadística básica; a continuación veremos qué significa esta última, ya que será importante para el futuro de este trabajo.

4.3. Área geoestadística básica

Para realizar los censos y las encuestas INEGI requirió definir las áreas geográficas de estudio. Esto los llevó a la creación del Marco Geoestadístico Nacional, el cual permite relacionar la información estadística con el espacio geográfico correspondiente. Este divide al territorio nacional en áreas de fácil identificación en campo y es adecuado para las actividades de captación de información. Estas divisiones geográficas se denominan Áreas Geoestadísticas y son: estatales (agee), municipales (agem) y básicas (ageb). Las ageb constituyen la unidad fundamental del Marco Geoestadístico. Dadas las diferencias de densidad de población y uso del suelo, se consideró necesario distinguir dos tipos de ageb: urbanas y rurales. Las ageb urbanas representan localidades con un número de habitantes mayor o igual a 2 500, o bien, una cabecera municipal independientemente de su número de pobladores, en conjuntos que generalmente van de 25 a 50 manzanas. Las ageb rurales enmarcan una superficie de aproximadamente 10 000 hectáreas, cuyo uso del suelo es predominantemente agropecuario y en ellas se encuentran distribuidas las localidades menores a 2 500 habitantes.

Una ventaja de estudiar a la población a nivel de ageb es que, por ser de mucho menor tamaño que el municipio e incluso que una localidad urbana, reducen los efectos de los grandes promedios estatales y municipales que tienden a suavizar y generalizar situaciones que evidentemente son diferentes.

Capítulo 5

Método para clasificación

En el capítulo anterior hemos presentado los estudios del nivel socioeconómico que se han hecho en México. Sin embargo, estos tienen dos inconvenientes. Primero, el INEGI realizó su estudió hace más de 20 años con los resultados del censo de población y vivienda del año 2000. Sin lugar a dudas, la situación social y económica de muchas regiones geográficas deben haber cambiado en el transcurso de ese tiempo; ya sea que hayan mejorado su situación o empeorado. Además, aunque se hubiese replicado el estudio con los resultados del censo del año 2010 (el censo más reciente) la diferencia de tiempo a la actualidad sigue siendo significativa, por lo que los resultados obtenidos pueden no estar apegados a la realidad.

La AMAI, por otro lado, al elaborar los estudios socioeconómicos con información de la encuesta enigh, proporciona resultados más actualizados puesto que se realiza con un periodo de dos años. La enigh no sólo proporciona datos sobre los ingresos y gastos de la población; también ofrece características ocupacionales y sociodemográficas de los integrantes de los hogares, así como las características de la infraestructura de la vivienda y el equipamiento del hogar. Esta encuesta puede ser una gran fuente información para obtener muchos de los indicadores del nivel socioeconómico, como los expuestos en la sección 4.1 del capítulo 4. No obstante, esta encuesta se elabora bajo un muestreo de hogares que permite presentar la información englobada a nivel entidad federativa. Es así que la AMAI presenta resultados del nivel socioeconómico a nivel estado y zona metropolitana, por lo cual, de requerir un análisis a una menor escala (como a nivel municipio, delegación o colonia) estos estudios no serían de mucha utilidad.

En nuestro trabajo, nos propusimos encontrar conjuntos de datos diferentes o complementarios a los que manejan el INEGI y la AMAI para poder realizar un estudio del nivel socioeconómico de las regiones del país, de manera que sea lo más apegado a la actualidad y a la menor escala que sea posible. Para ello, nos servimos de las herramientas y algoritmos de aprendizaje automático para clasificar a las regiones del país de acuerdo a sus características sociales y económicas.

A continuación, veremos cuáles fueron los conjuntos de datos que utilizamos y cómo es que fueron elegidos. Posteriormente explicaremos qué métodos de aprendizaje automático utilizamos y cómo hicimos la clasificación de las regiones.

5.1. Descripción de los conjuntos de datos

El INEGI tiene un amplio catálogo de censos y encuestas cuyos resultados son presentados públicamente en su sitio web. Los que econtramos que pudieran ser útiles para un estudio socioeconómico fueron:

1. Censo de Población y Vivienda 2010.
2. Encuesta Nacional de Ingresos y Gastos de los Hogares (enigh) 2016.
3. Encuesta Nacional sobre Disponibilidad y Uso de Tecnologías de la Información en Hogares (endutih) 2017.
4. Encuesta Nacional en Hogares (enh) 2017.
5. Encuesta Nacional de Ocupación y Empleo (enoe) 2018.
6. Características del Entorno Urbano 2014.
7. Directorio Estadístico Nacional de Unidades Económicas (denue).

Sin embargo, como se mencionó previamente para la encuesta enigh, las encuestas endutih, enh y enoe también son muestrales, lo que significa que no son las adecuadas para nuestro estudio. A continuación describiremos un poco al conjunto de datos del censo de población y vivienda 2010 y al directorio estadístico nacional de unidades económicas (denue).

5.1.1. Censo de población y Vivienda

Los censos de población y vivienda constituyen la fuente de información estadística más completa de la población del país. En el censo del 2010, esta información se puede encontrar hasta un nivel de área geoestadísica básica (ageb) y manzana. Los indicadores incluidos para población incluyen apectos como sexo, edad, fecundidad, migración, lengua indígena, discapacidad, características educativas, características económicas, derechohabiencia a servicios de salud, situación conyugal y religión. En cuanto a viviendas, se ofrece información sobre: total de viviendas y ocupantes, material de pisos, número de cuartos, servicios disponibles en la vivienda (energía eléctrica, agua entubada, sanitario, drenaje) así como los bienes con los que cuenta la misma. El conjunto de datos que proporciona el INEGI en su página web para este censo, contiene 190 indicadores obtenidos de los resultados para 56 195 agebs que conforman las localidades urbanas del país [INEGI, a].

5.1.2. Directorio de Unidades Económicas

El directorio estadístico nacional de unidades económicas, como su nombre lo indica, es un conjunto de datos conformado por un registro de negocios y establecimientos cuyo propósito es proveer los datos de identificación, ubicación y contacto de las unidades económicas a usuarios tanto especializados como no especializados [INEGI, c]. El denue proporciona información de las unidades económicas cuyas actividades están ordenadas con base en el Sistema de Clasificación Industrial de América del Norte (SCIAN México) [INEGI, b]. Los códigos de actividades el SCIAN están agrupados como se muestra en la figura 5.1, donde puede observarse que estos están agrupados en 5 niveles. El primer nivel está compuesto por códigos de 2 dígitos, el segundo de 3, y así sucesivamente.

43	***Comercio al por mayor*** ^{T}
431	**Comercio al por mayor de abarrotes, alimentos, bebidas, hielo y tabaco**
4311	*Comercio al por mayor de abarrotes y alimentos*
43111	*Comercio al por mayor de abarrotes*
431110	Comercio al por mayor de abarrotes
43112	*Comercio al por mayor de carnes*
431121	Comercio al por mayor de carnes rojas
431122	Comercio al por mayor de carne de aves
431123	Comercio al por mayor de pescados y mariscos

Figura 5.1: Organización de los códigos de actividades económicas del SCIAN.

5.2. Descripción del procedimiento

Considerando que las empresas e instituciones no se expanden en cualquier sitio, sino en lugares donde las condiciones sociales y económicas de los habitantes se adaptan a sus necesidades; nuestro objetivo es construir un modelo de aprendizaje supervisado que clasifique a las agebs del país de acuerdo a las unidades económicas que haya en ellas. Para ello, hacemos uso del conjunto de datos del denue. No obstante, para entrenar dicho modelo requerimos primero de un conjunto de datos de aprendizaje. Similar a lo que hizo INEGI en el 2004 ([INEGI, 2004]), podemos realizar un agrupamiento de las agebs con la información del censo de población y vivienda 2010; estos resultados pueden ser utilizados para crear un conjunto de prueba para así entrenar el modelo de clasificación. Por lo tanto, nuestro trabajo consiste en dos etapas:

1. Agrupamiento inicial de agebs utilizando la información del censo de población y vivienda 2010.

2. Creación de un modelo de clasificación de agebs con los datos del denue, empleando los resultados del agrupamiento para entrenamiento.

Se eligió clasificar a nivel de agebs debido a que el censo del 2010 maneja la información a este nivel, además de que el denue también especifica la ageb a la que pertece cada unidad económica. En la sección siguiente se explica cómo fue el procedimiento llevado a cabo para la primera etapa de nuestro trabajo.

5.3. Agrupamiento de agebs

Como primer paso, necesitamos crear un conjunto de indicadores a partir de las variables que existen en el conjunto de datos de resultados del censo del año 2010. Estas nuevas variables deben proporcionar información que describan características socioeconómicas de la población. Para esto, tomamos en cuenta los indicadores que propuso INEGI en el 2004 (apéndice A). En el conjunto de datos de los resultados del censo del 2010 no hay información respecto a salarios de los habitantes ni al hacinamiento de las viviendas, entre algunos otros, por lo que no fue posible construir todos estos indicadores; aunque sí construimos algunos nuevos. Los indicadores que inicialmente propusimos se muestran en la tabla 5.1. En contraste con los indicadores del apéndice A, estos nuevos incluyen la tenencia de computadora e internet, lo que evidencia el hecho de que en el 2000 estos indicadores aún no eran factores relevantes como en el 2010.

IND01	Porcentaje de viviendas particulares con agua entubada
IND02	Porcentaje de viviendas particulares con energía eléctrica
IND03	Porcentaje de viviendas particulares con drenaje
IND04	Porcentaje de viviendas particulares piso diferente de tierra
IND05	Porcentaje de viviendas particulares con servicio sanitario
IND06	Porcentaje de viviendas particulares con refrigerador
IND07	Porcentaje de viviendas particulares con televisión
IND08	Porcentaje de viviendas particulares con teléfono fijo
IND09	Porcentaje de viviendas particulares con teléfono móvil
IND10	Porcentaje de viviendas particulares con internet
IND11	Porcentaje de viviendas particulares con computadora
IND12	Porcentaje de viviendas particulares con lavadora
IND13	Porcentaje de viviendas particulares con automóvil o camioneta propios
IND14	Porcentaje de población con derechohabiencia a servicios de salud
IND15	Porcentaje de población de 15 años y más alfabeta
IND16	Porcentaje de población de 6 a 11 años que asisten a la escuela
IND17	Porcentaje de población de 12 a 14 años que asisten a la escuela
IND18	Porcentaje de población de 15 a 17 años que asisten a la escuela
IND19	Porcentaje de población de 15 años y más con secundaria completa
IND20	Porcentaje de población de 18 años y más con educación pos-básica
IND21	Porcentaje de población ocupada femenina
IND22	Porcentaje de población económicamente activa

Tabla 5.1: Indicadores construidos con los resultados del censo de población y vivienda 2010

Se construyó un nuevo conjunto de datos que capturan la información de estos 22 indicadores por cada ageb de México. Cabe mencionar que las nuevas variables están construidas de manera que a mayor porcentaje, mejor situación. Las fórmulas para

obtenerlos se describen en el apéndice B.

Como parte del preprocesamiento del nuevo conjunto de datos se eliminaron aquellas agebs con población cero o que tuvieran valores faltantes, con el propósito de prevenir errores futuros. Posteriormente, se aplicó un análisis de correlación y un proceso de detección de valores atípicos. Al realizar el análisis de correlación los indicadores 10 y 11 presesentaron una correlación de 0,97, así que es redundante conservar las dos variables, por lo que se descarta el indicador 11 (porcentaje de viviendas particulares con computadora). La matriz de correlación queda ilustrada en la figura 5.2.

Como método de detección de valores atípicos, se utilizó el algoritmo basado en densidad RKOF (Robust Kernel-based Outlier Factor) [Gao et al., 2011] para eliminar el 2 % de los datos de acuerdo a los puntajes atípicos. En la figura 5.3 se muestran dos gráficas, una antes y otra después de remover los valores atípicos.

Con lo anterior, se tiene un conjunto de datos de 43 990 agebs con 21 atributos. Como algoritmo de agrupamiento se utilizó k-means, ya que este es uno de los más utilizados para datos censales y geogemográficos [Harris et al., 2005]; también se utilizó el algoritmo clara. Estos algoritmos tienen además la ventaja de que, al ser basados en prototipos, permiten ordenar fácilmente a los grupos resultantes mediante el método descrito en la sección 4.2 del capítulo 4.

Al igual que el estudio de INEGI en el 2004, hicimos un agrupamiento de las agebs con un número de grupos $k = 7$. No obstante, también optamos por realizar agrupamientos para $k = 5$ y $k = 3$. Las gráficas de las agebs agrupadas con k-means respecto a las dos primeras componentes principales para $k = 7$, $k = 5$ y $k = 3$, se muestran en las figuras 5.4, 5.6 y 5.8, respectivamente; mientras que los agrupamientos hechos con clara están graficados en las figuras 5.5, 5.7 y 5.9.

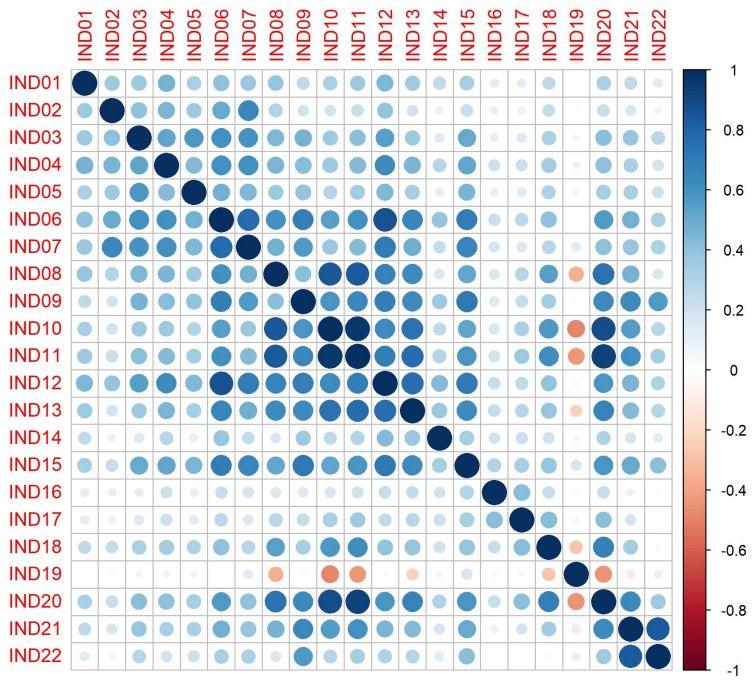

Figura 5.2: Matriz de correlación para el conjunto de datos de indicadores.

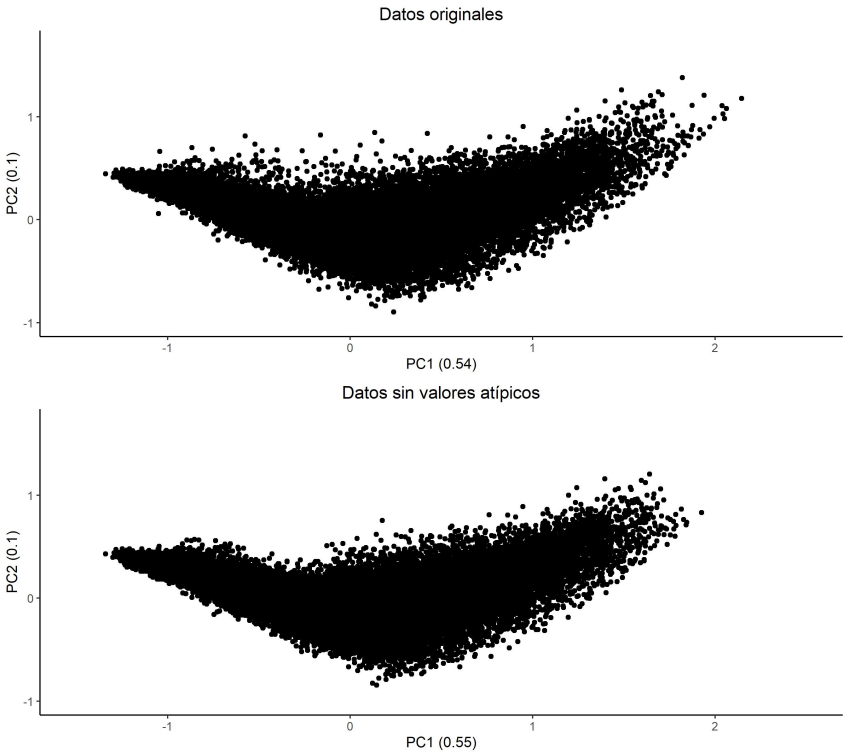

Figura 5.3: Comparación de las gráficas de los datos antes y después de aplicarles detección de valores atípicos.

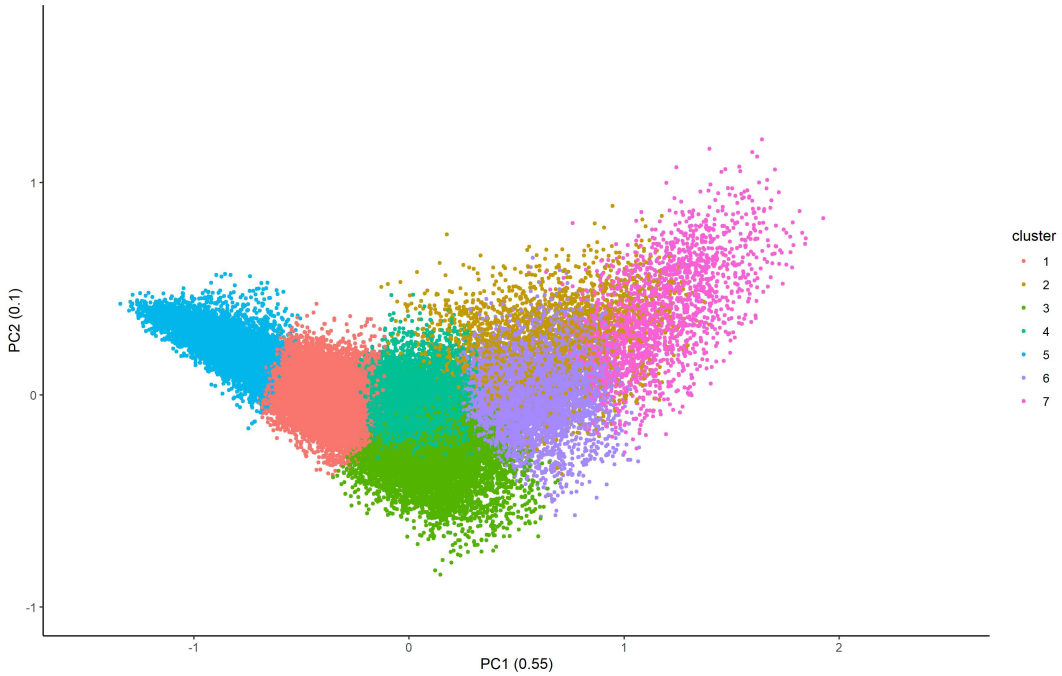

Figura 5.4: Agrupamiento por k-means para $k = 7$.

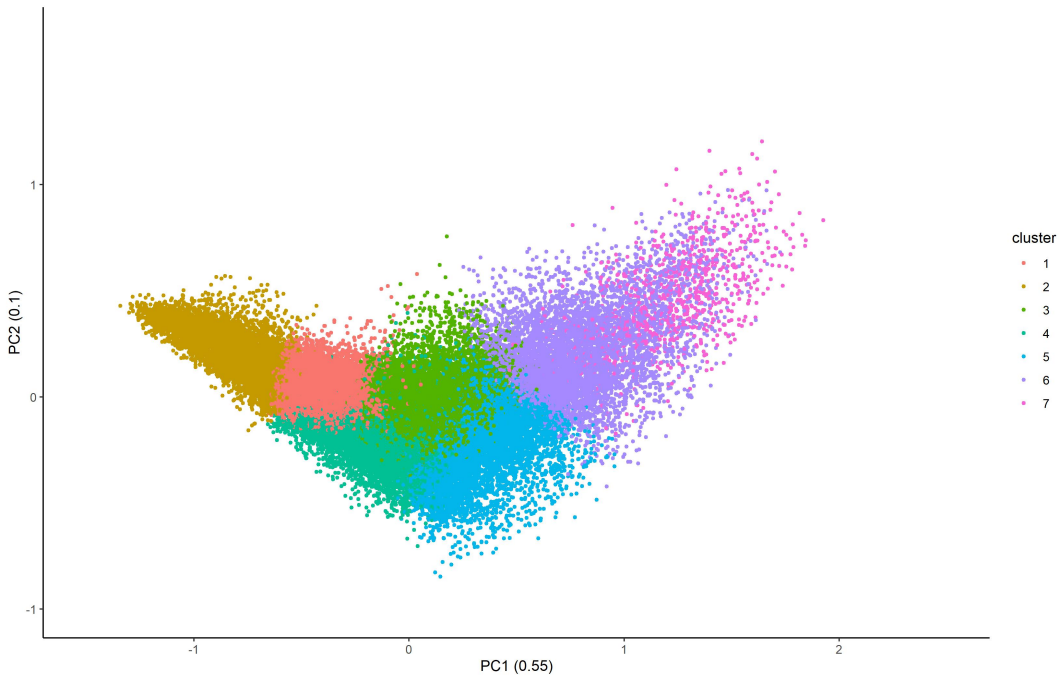

Figura 5.5: Agrupamiento por clara para $k = 7$.

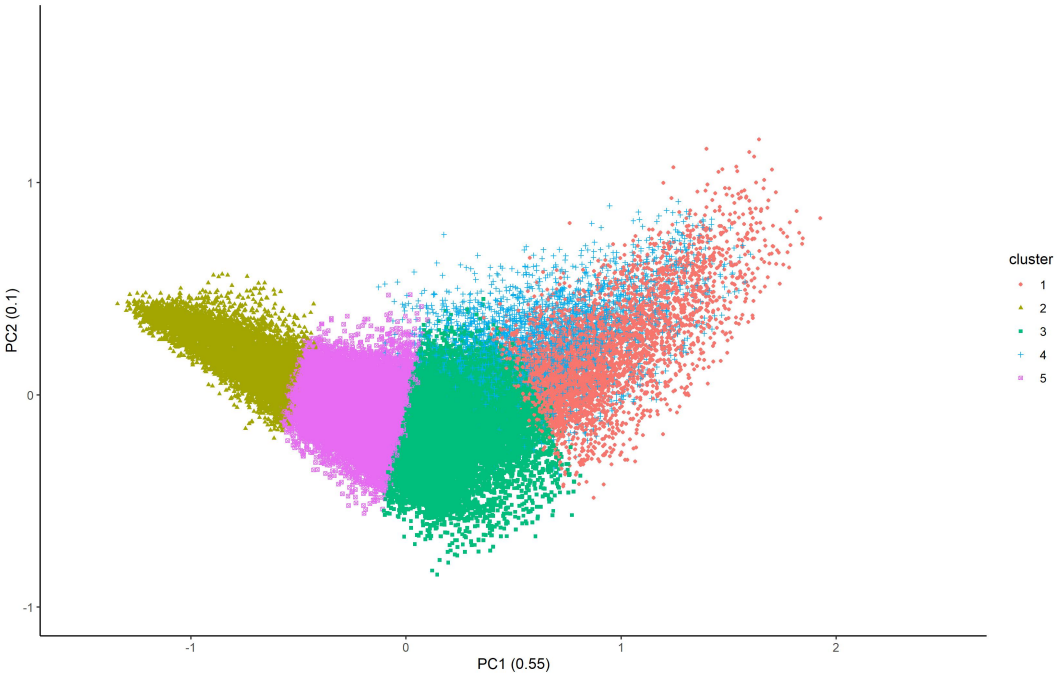

Figura 5.6: Agrupamiento por k-means para $k = 5$.

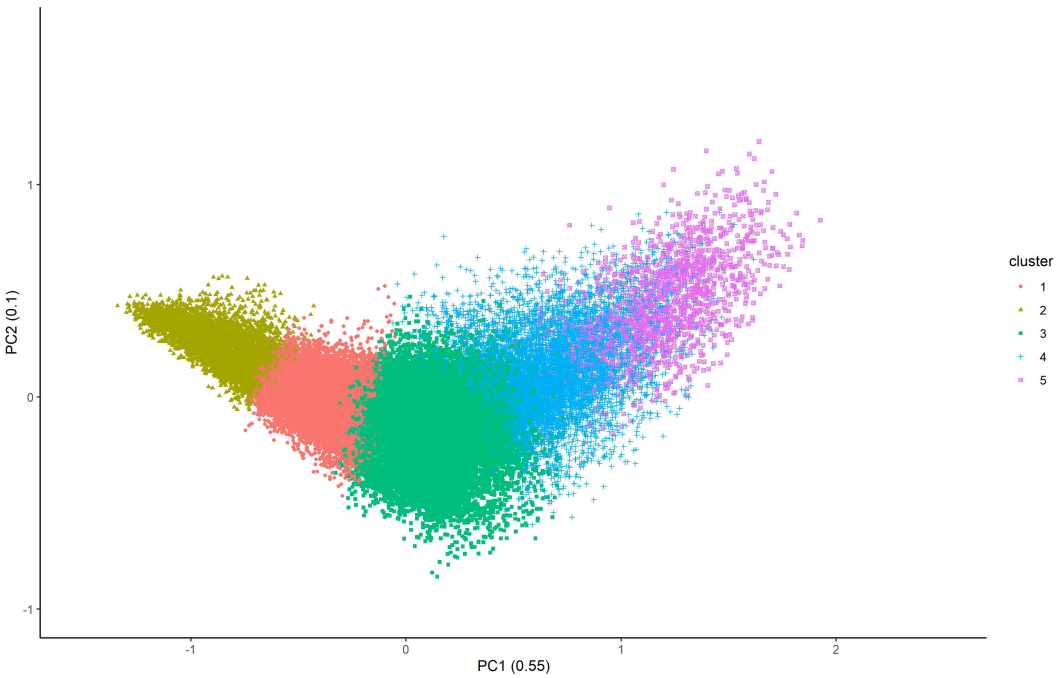

Figura 5.7: Agrupamiento por clara para $k = 5$.

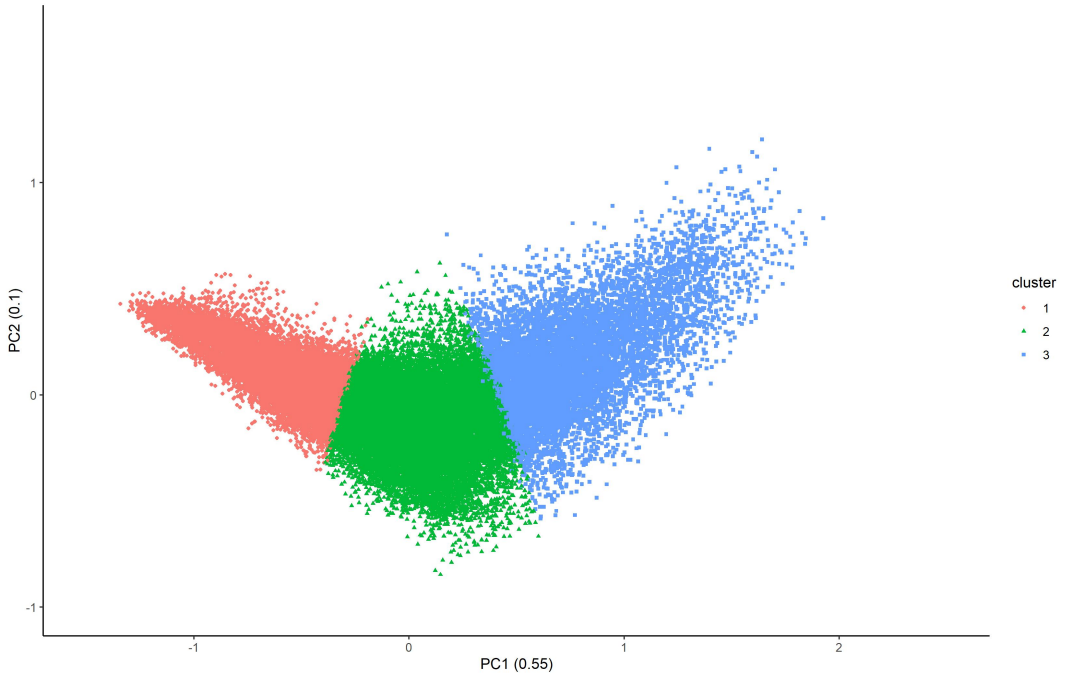

Figura 5.8: Agrupamiento por k-means para $k = 3$.

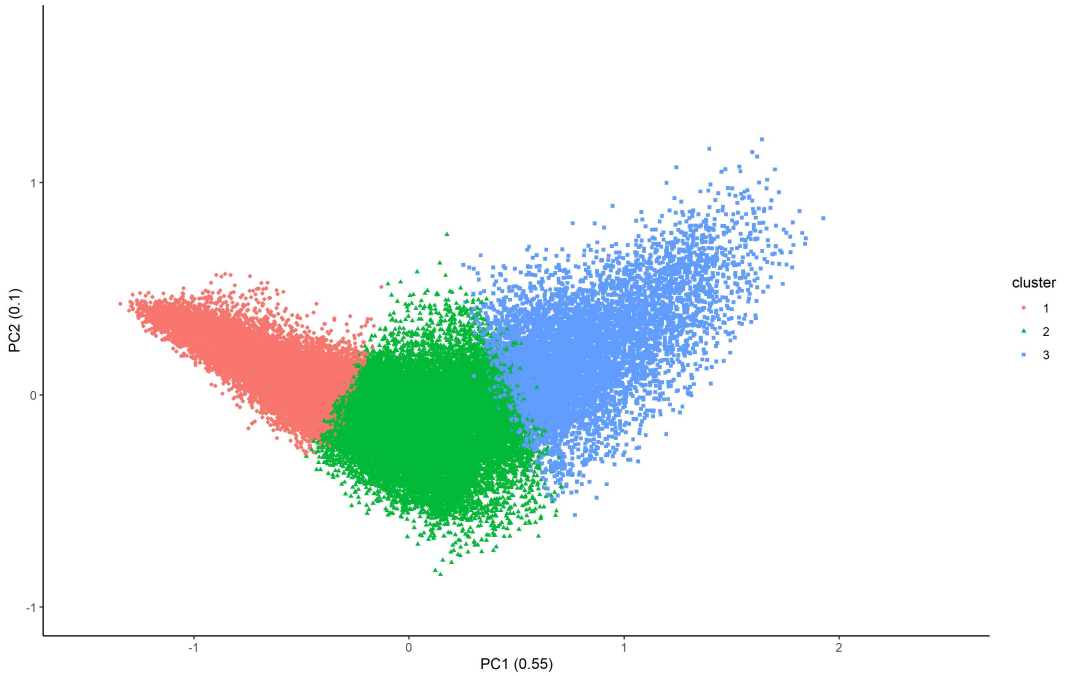

Figura 5.9: Agrupamiento por clara para $k = 3$.

Mediante el método de ordenamiento ya mencionado, es decir, mediante las distancias entre los centros, se ordenaron los grupos de manera que el grupo que representa

a las agebs con la mejor situación socioeconómica relativa se identificara como estrato 7 y el más desfavorecido como estrato 1 (para el caso $k = 7$). Lo mismo aplicó para los agrupamientos con $k = 5$ y $k = 3$. Las correspondencias de los grupos con los estratos adecuados se muestran en las tablas 5.2, 5.3 y 5.4 para 7, 5 y 3 grupos, respectivamente.

k-means $k = 7$		
Grupo	Estrato	No. de agebs
5	7	4993
1	6	10671
4	5	10478
3	4	6952
6	3	6166
2	2	2732
7	1	1998

clara $k = 7$		
Grupo	Estrato	No. de agebs
2	7	5430
1	6	7780
4	5	6649
3	4	11481
5	3	6222
6	2	5564
7	1	864

Tabla 5.2: Relación de los grupos formados para $k = 7$, con los estratos de nivel socioeconómico correspondientes.

k-means $k = 5$		
Grupo	Estrato	No. de agebs
2	5	6864
5	4	15476
3	3	14507
4	2	3005
1	1	4138

clara $k = 5$		
Grupo	Estrato	No. de agebs
2	5	4900
1	4	10705
3	3	19760
4	2	7225
5	1	1400

Tabla 5.3: Relación de los grupos formados para $k = 5$, con los estratos de nivel socioeconómico correspondientes.

k-means $k = 3$		
Grupo	Estrato	No. de agebs
1	3	11911
2	2	23434
3	1	8645

clara $k = 3$		
Grupo	Estrato	No. de agebs
1	3	11853
2	2	24623
3	1	7514

Tabla 5.4: Relación de los grupos formados para $k = 3$, con los estratos de nivel socioeconómico correspondientes.

Estas tablas también contienen la distribución de agebs por estratos. Para una mejor apreciación, las gráficas de distrubución de agebs en el país se presentan en las figuras 5.10 a 5.15. Las gráficas de distrubución de agebs por estado se muestran en el anexo C.

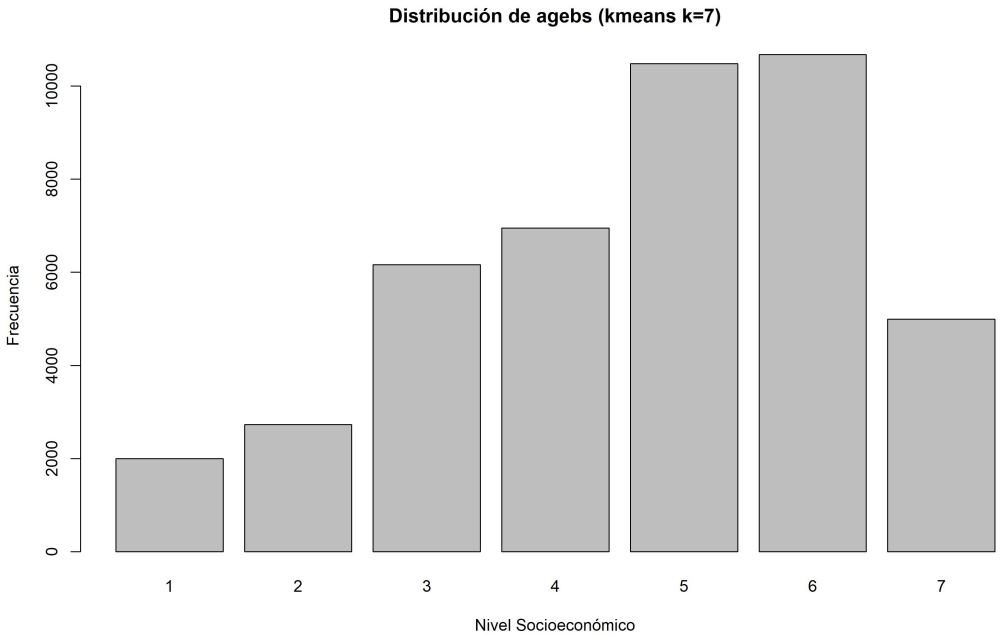

Figura 5.10: Distribución de agebs para $k = 7$ (k-means).

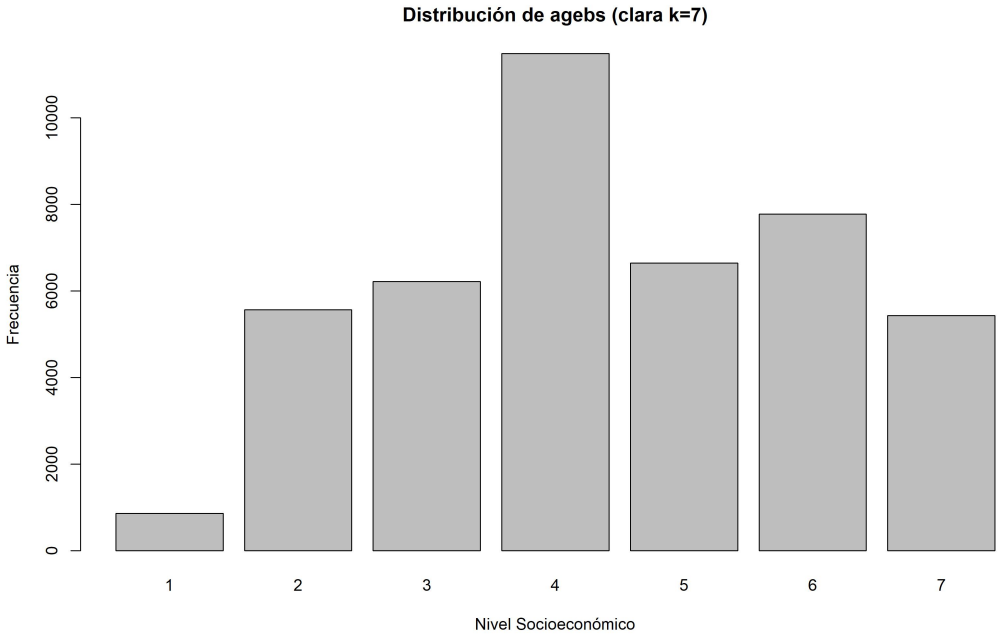

Figura 5.11: Distribución de agebs para $k = 7$ (clara).

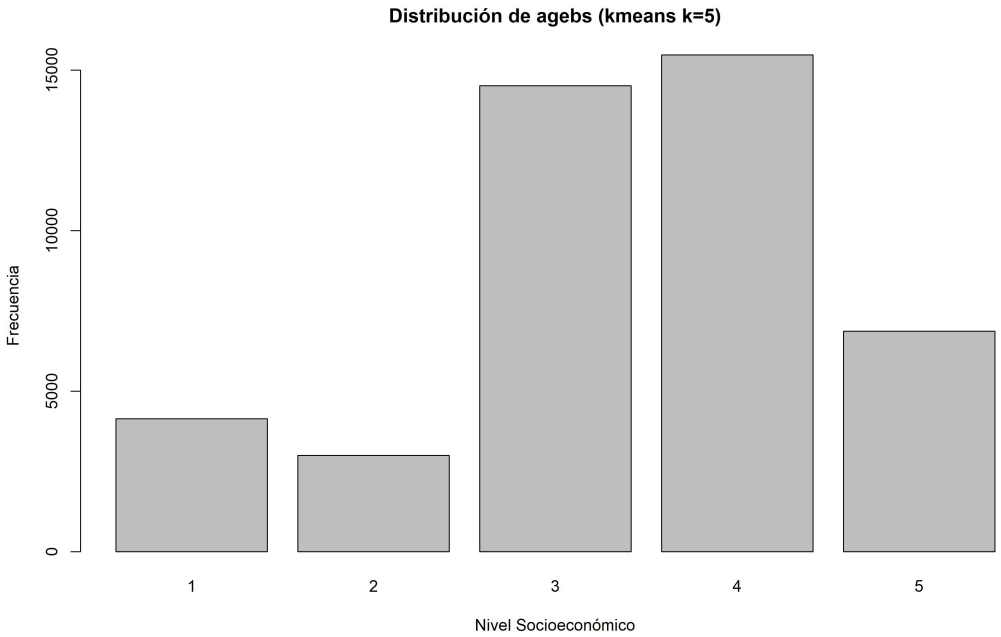

Figura 5.12: Distribución de agebs para $k = 5$ (k-means).

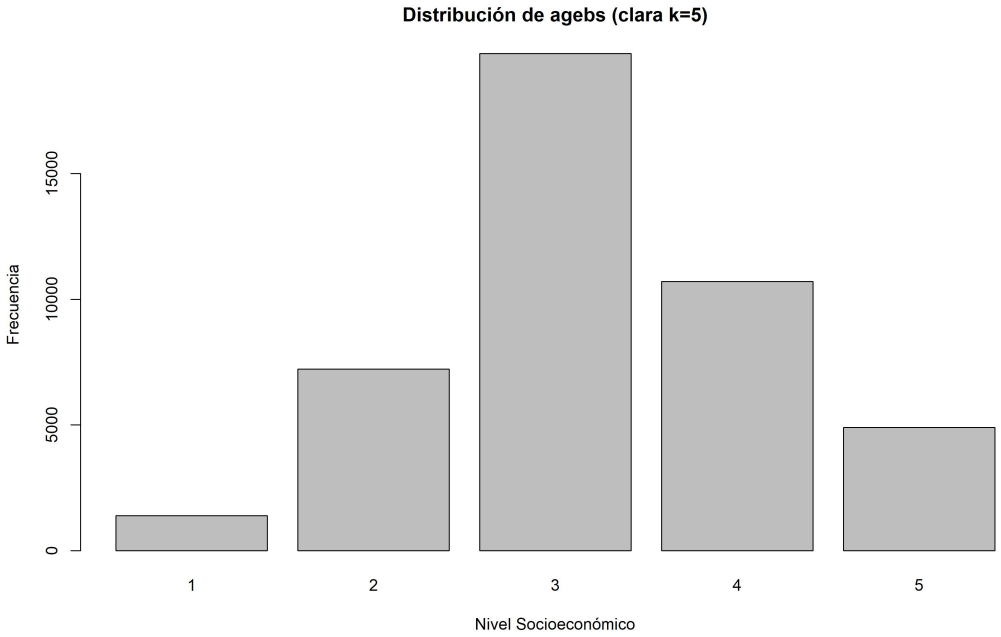

Figura 5.13: Distribución de agebs para $k = 5$ (clara).

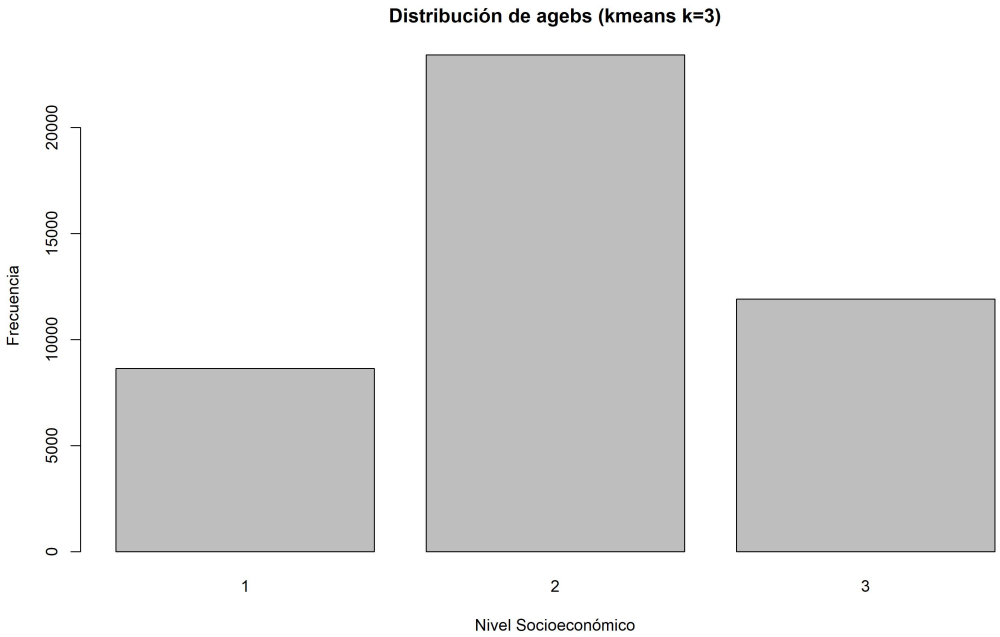

Figura 5.14: Distribución de agebs para $k = 3$ (k-means).

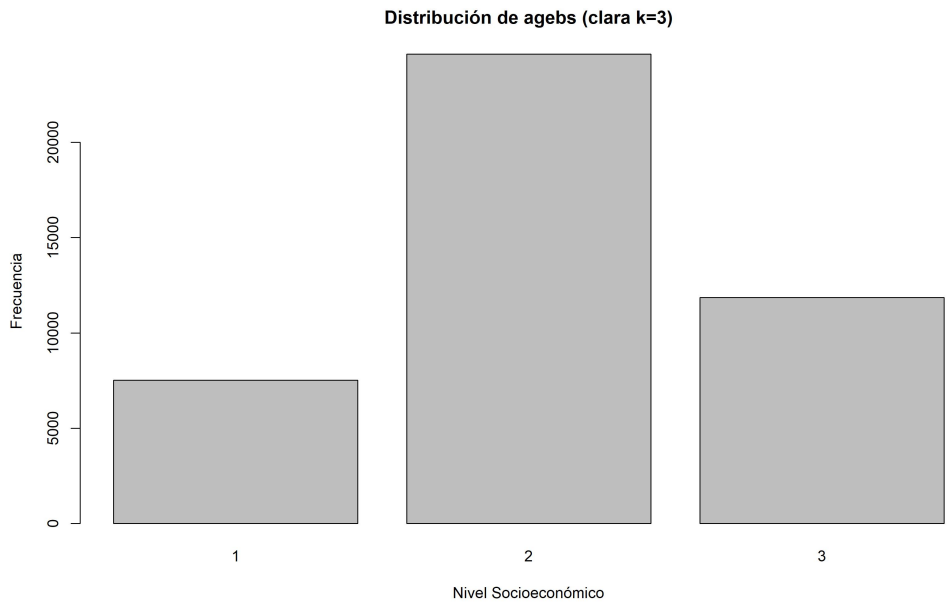

Figura 5.15: Distribución de agebs para $k = 3$ (clara).

Influencia	Indicador
100.00 %	IND01
100.00 %	IND10
100.00 %	IND12
99.99 %	IND06
97.81 %	IND13
97.77 %	IND20
97.20 %	IND08
95.81 %	IND09
90.12 %	IND18
85.62 %	IND14
79.66 %	IND19
77.21 %	IND15
74.55 %	IND21
72.29 %	IND22
69.50 %	IND03
64.62 %	IND07
61.63 %	IND04
58.60 %	IND05
48.95 %	IND16
48.67 %	IND02
40.90 %	IND17

Tabla 5.5: Influencia de los indicadores utilizados para el agrupamiento.

Posterior al agrupamiento, se aplicó un árbol de decisión sobre los datos como un proceso de selección de características. De esta manera, se pueden desechar aquellas variables que tengan poca influencia sobre los resultados finales. La influencia de cada uno de los indicadores utilizados para el agrupamiento se muestra en la tabla 5.5. Como puede observarse, todas las variables resultaron importantes.

5.4. Modelo de clasificación

Como se mencionó anteriormente, el objetivo es crear un modelo de clasficación de agebs a partir de la información del denue. Sin embargo, este es un directorio, por lo que necesitamos crear un nuevo conjunto de datos a partir de él. Lo que se hizo fue un conteo de cuántas unidades económicas y de qué tipo hay para cada ageb, de manera que los atributos del conjunto de datos sean el tipo de unidad económica. Este conteo se elaboró sobre los códigos de nivel 2 de SCIAN, es decir, de 3 dígitos; de tomar niveles más altos (códigos con más digitos) crecería demasiado el número de atributos, lo que complicaría el entrenamiento del modelo de clasificación.

Con este nuevo conjunto de datos, se procedió a observar las variables para determinar si en realidad la distribución de los tipos de unidades económicas iba acorde con el nivel socioeconómico del ageb. Este análisis de variables fue revelador, ya que, como se puede constratar en las gráficas de las figuras 5.16 y 5.17, las escuelas primarias privadas se encuentran predominantemente en los niveles socioeconómicos más altos mientras que en las públicas no hay una distinción clara. Así como con las primarias públicas, el mismo comportamiento se puede observar en otros tipos de unidades económicas. Fue en base a esta observación de variables que se eliminaron algunos tipos de unidades económicas que no presentaran una distinción por niveles socioeconómicos, y que por tanto proporcionara información no relevante o contraproducente para el modelo. En total fueron 227 los tipos de unidades eliminados.

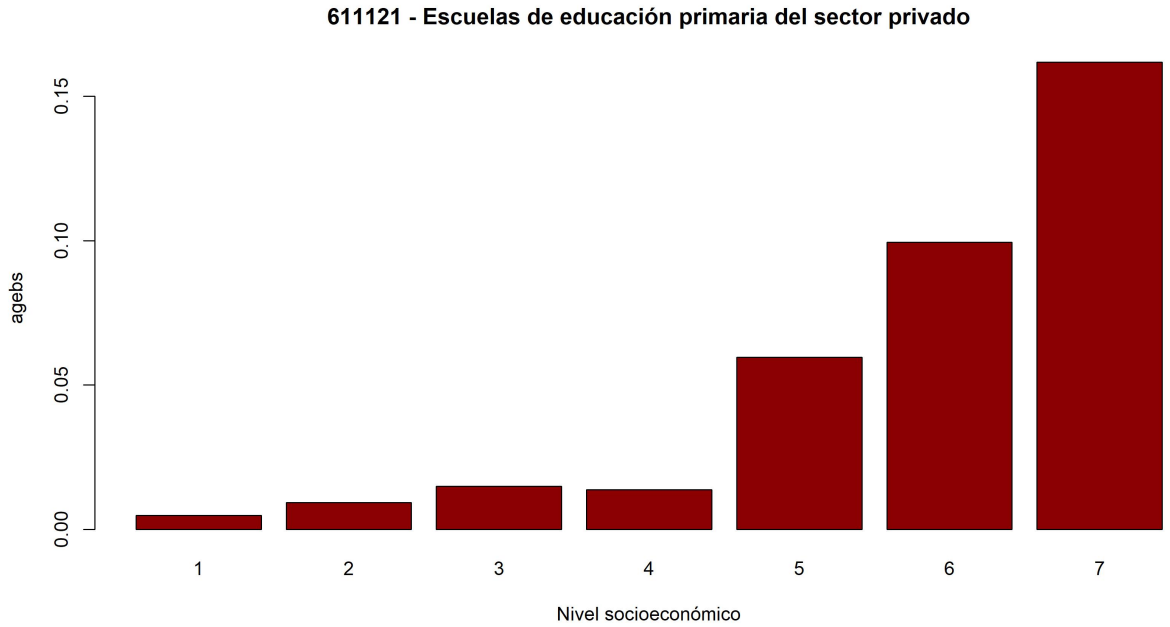

Figura 5.16: Distribución de agebs con escuelas primarias del sector privado por nivel socioeconómico.

Al final se trabajó sobre un conjunto de datos con 80 atributos (tipos de unidades económicas) para cada una de las agebs, de un total de 40 644. Esta cantidad es menor respecto a las manejadas en la parte de agrupamiento con los datos del censo, debido a que hay algunas agebs que no poseen unidades económicas de ningún tipo, o bien, agebs que no se tomaron en cuenta en la primera etapa de nuestro trabajo. Después de normalizar los datos, se probaron los algoritmos de clasficiación: álbol de decisión, red neuronal, máquina de soporte vectorial, bosque aleatorio y un método de boosting. Los resultados se muestran en el cápítulo siguiente.

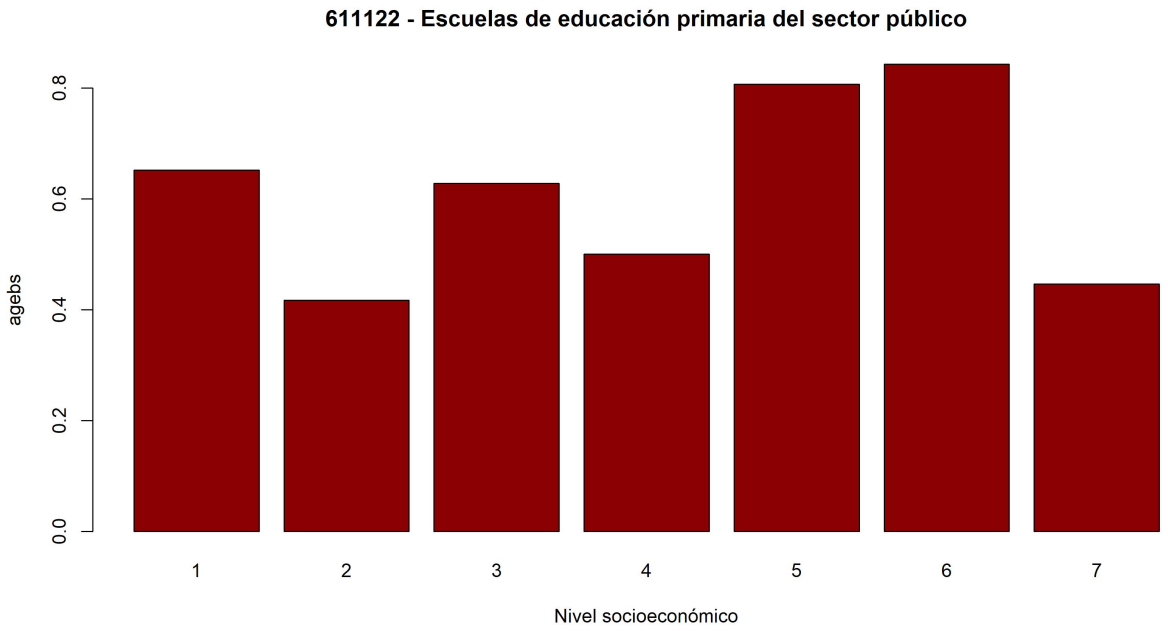

Figura 5.17: Distribución de agebs con escuelas primarias del sector público por nivel socioeconómico.

Capítulo 6

Resultados

Sobre el conjunto de datos de unidades económicas por ageb,se aplicaron implementaciones en el leguaje de programación R de los algoritmos: árbol de decisión (C5.0), red neuronal (rna), máquina de soporte vectorial (svm), bosque aleatorio (rf) y un método de boosting (gbm). La idea es emplear como método de clasificación final para nuestro modelo a aquel que otorgue la mejor exactitud. Se probó tanto el agrupamiento inicial hecho por k-means como el hecho por clara, y también para los tres valores de k (7, 5 y 3). El conjunto de datos se separó de manera aleatoria en un 70 % para un conjunto de entrenamiento y el 30 % restante para conjunto de prueba. La exactitud obtenida por las pruebas de clasificación se muestran en la tabla 6.1 para el agrupamiento inicial hecho por k-means, y en la tabla 6.2 para el agrupamiento con clara.

Exactitud			
	$k = 7$	$k = 5$	$k = 3$
C5.0	0.404	0.525	0.663
svm	0.431	0.539	0.652
rf	0.464	0.559	0.684
gbm	0.461	0.557	0.682
rna	0.454	0.555	0.676

Tabla 6.1: Exactitud obtenida por cada algoritmo de clasificación y para cada número de clases. Agrupamiento inicial con k-means.

Como se puede apreciar, las clasificaciones para un número de $k = 7$ clases resultan con una mejor exactitud para el agrupamiento con k-means, mientras que para 3 y 5 clases, el agrupamiento con clara resultó mejor. Hablando de los algoritmos de clasificación, en todos se mantiene el orden $rf > gbm > rna > svm > C5{,}0$. El bosque aleatorio es el más eficiente en términos de exactitud en todos los casos, con una ligera ventaja sobre el algoritmo de boosting, gbm. Naturalmente, el menos eficaz fue el árbol de decisión C5.0, dado que es el más sencillo de todos.

También podemos notar que a medida que disminuimos el número de clases, obtenemos mejores clasificaciones. Esto es entendible ya que las clases no son simples

Exactitud			
	$k = 7$	$k = 5$	$k = 3$
C5.0	0.408	0.553	0.690
svm	0.430	0.557	0.680
rf	0.457	0.585	0.703
gbm	0.455	0.581	0.702
rna	0.447	0.580	0.696

Tabla 6.2: Exactitud obtenida por cada algoritmo de clasificación y para cada número de clases. Agrupamiento inicial con clara.

etiquetas, sino niveles; entonces la clase 7 tiene más en común con las clases 6 y 5 que con las clases 1 y 2, para el caso $k = 7$. El mejor resultado lo obtenemos del bosque aleatorio con el agrupamiento de clara, con una exactitud del 70.3 %.

En las tablas siguientes se muestran las medidas de evaluación de recuerdo, precisión y F-score para las clasificaciones realizadas con 3 clases. Las tablas de estas medidas para el resto de clasificaciones están presentes en el apéndice D.

Medidas de evaluación de rf con k-means $k = 3$			
clase	prec	rec	F
1	0.6888412	0.3008435	0.4187867
2	0.6611468	0.8807023	0.7552925
3	0.7680369	0.5393225	0.6336736

Tabla 6.3: Medidas de evaluación para el bosque aleatorio con 3 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rf con clara $k = 3$			
clase	prec	rec	F
1	0.7485265	0.2087671	0.3264781
2	0.6816672	0.9093768	0.7792271
3	0.7766908	0.5415184	0.6381269

Tabla 6.4: Medidas de evaluación para el bosque aleatorio con 3 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de gbm con k-means $k = 3$			
clase	prec	rec	F
1	0.6657116	0.3266167	0.4382270
2	0.6633345	0.8662965	0.7513503
3	0.7568563	0.5446244	0.6334361

Tabla 6.5: Medidas de evaluación para el algoritmo de boosting con 3 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de gbm con clara $k = 3$			
clase	prec	rec	F
1	0.6793785	0.2635616	0.3797868
2	0.6876791	0.8919383	0.7766024
3	0.7685608	0.5495255	0.6408439

Tabla 6.6: Medidas de evaluación para el algoritmo de boosting con 3 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de rna con k-means $k = 3$			
clase	prec	rec	F
1	0.6483051	0.2867854	0.3976608
2	0.6608153	0.8586435	0.7468511
3	0.7401544	0.5646539	0.6406015

Tabla 6.7: Medidas de evaluación para la red neuronal con 3 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rna con clara $k = 3$			
clase	prec	rec	F
1	0.6818182	0.2054795	0.3157895
2	0.6850595	0.8802173	0.7704723
3	0.7362472	0.5794781	0.6485231

Tabla 6.8: Medidas de evaluación para la red neuronal con 3 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de svm con k-means $k = 3$			
clase	prec	rec	F
1	0.8917526	0.08106842	0.1486254
2	0.6255895	0.91566627	0.7433305
3	0.7487751	0.49513991	0.5960993

Tabla 6.9: Medidas de evaluación para la máquina de soporte vectorial con 3 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de svm con clara $k = 3$			
clase	prec	rec	F
1	0.8479532	0.07945205	0.1452906
2	0.6590653	0.92124071	0.7684054
3	0.7583593	0.50444840	0.6058771

Tabla 6.10: Medidas de evaluación para la máquina de soporte vectorial con 3 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de C5.0 con k-means $k = 3$			
clase	prec	rec	F
1	0.6223776	0.3336457	0.4344112
2	0.6578037	0.8278812	0.7331074
3	0.6979715	0.5472754	0.6135050

Tabla 6.11: Medidas de evaluación para el árbol de decisión con 3 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de C5.0 con clara $k = 3$			
clase	prec	rec	F
1	0.6507937	0.2920548	0.4031770
2	0.6885662	0.8539165	0.7623788
3	0.7075612	0.5661329	0.6289951

Tabla 6.12: Medidas de evaluación para el árbol de decisión con 3 clases. Agrupamiento inicial hecho con clara.

Como se mencionó en el capítulo 3, los algoritmos como el árbol de decisión y el bosque aleatorio pueden proporcionar información sobre la influencia o importancia de los atributos. En nuestras pruebas, los tres tipos de unidades económicas con mayor importancia en la mayoría de los casos fueron los servicios educativos (código 611), los servicios de reparación y mantenimiento (código 811) y el comercio al por menor de

artículos de papelería, para el esparcimiento y otros artículos de uso temporal (código 465). Es decir, estas tres variables son las que más se utilizaron para la clasificación por nivel socioeconómico de las agebs. Sus gráficas de distribución por ageb se muestran en las figuras 6.1, 6.2 y 6.3.

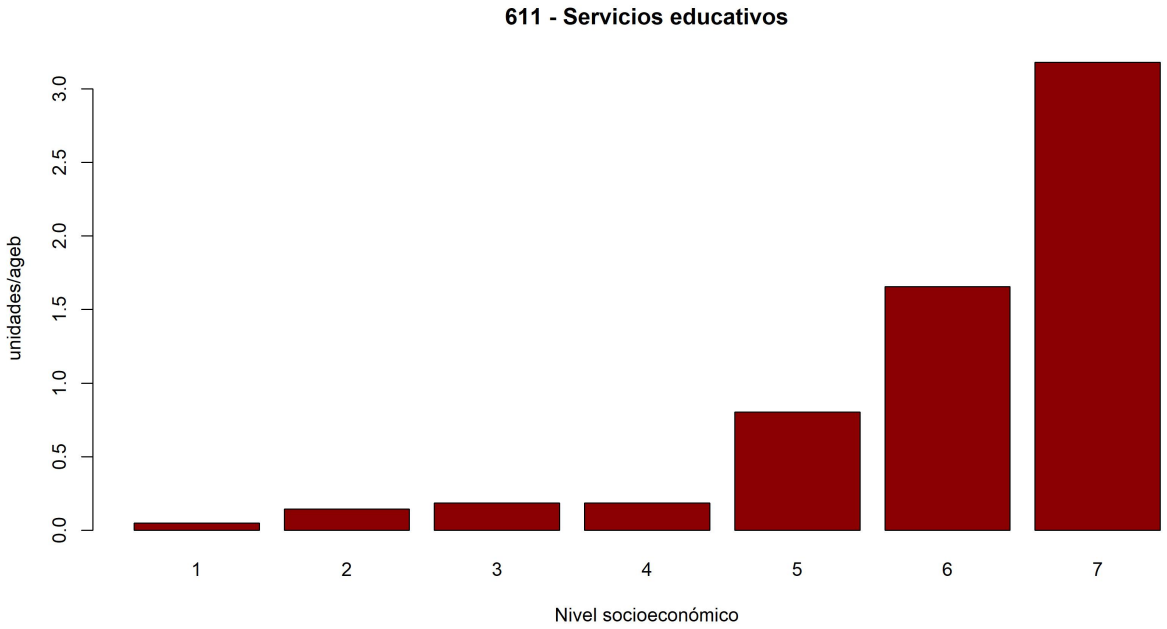

Figura 6.1: Distribución de agebs con servicios educativos por nivel socioeconómico.

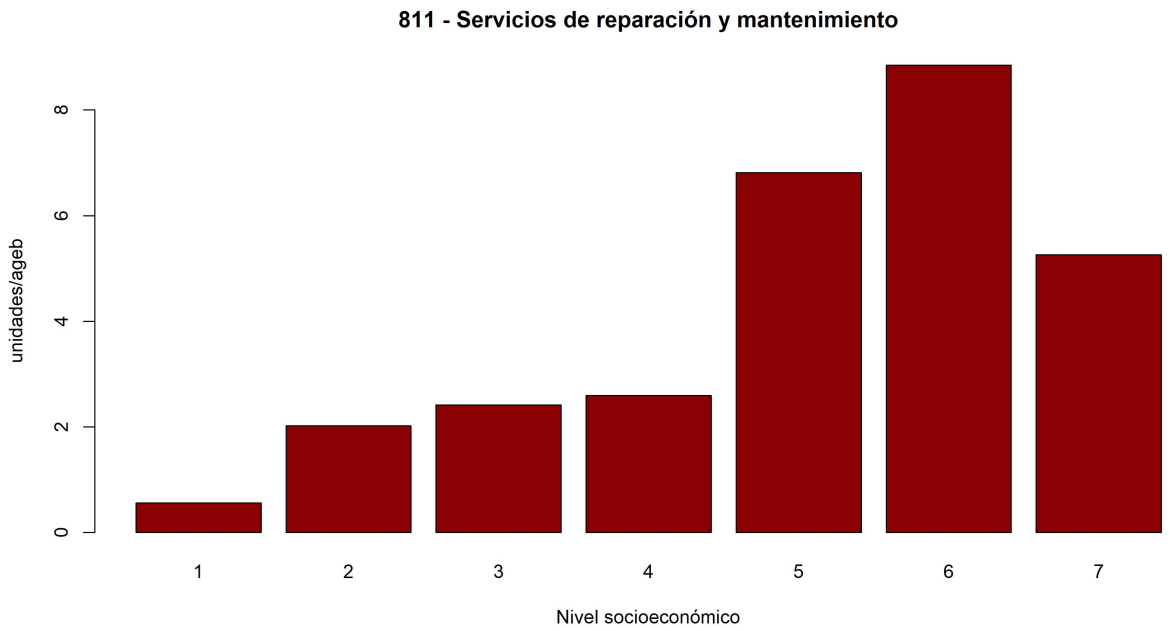

Figura 6.2: Distribución de agebs con servicios de reparación y mantenimiento por nivel socioeconómico.

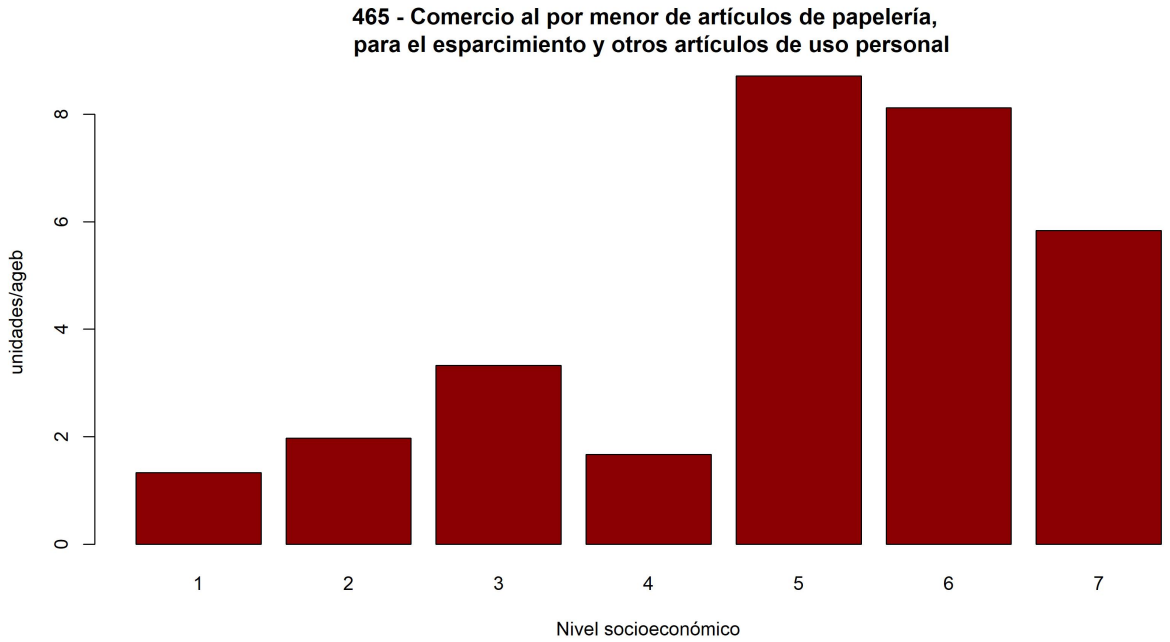

Figura 6.3: Distribución de agebs con comercio al por menor de artículos de papelería, para el esparcimiento y otros artículos de uso temporal por nivel socioeconómico.

Capítulo 7

Conclusiones

Como se puede observar de los resultados, el modelo de clasificación creado por el bosque aleatorio resultó ser el más efectivo en términos de exactitud en todos los casos, llegando hasta un 70.3 % para la clasificación con 3 clases y con el agrupamiento inicial realizado por el algoritmo clara. En cuanto a las clases, la clase 2 (la que representa a las agebs con el nivel socioeconómico intermedio) fue en la que se obtuvo un mejor F-score con 0.779, mientras que la clase 1 (la más baja) tuvo un F-score de 0.326.

Si nos quedamos con la clasificación en 7 clases, como lo manejó el INEGI en su estudio del 2004, la mejor exactitud conseguida fue de 46.4 %; exactitud obtenida mediante el bosque aleatorio y con el agrupamiento inicial de k-means. Estos resultados no parecen prometedores, aunque bien se pueden reducir el número de clases para mejorarlos.

Muchos tipos de unidades económicas presentan una distribución predominante en determinados estratos socioeconómicos, como las escuelas privadas que predominan en los niveles altos. No obstante, es probable que el mero conocimiento del tipo y número de unidades económicas que hay en un región no sea información suficiente para determinar su nivel socioeconómico. Sin embargo también hay algunos otros factores que se pueden tomar en cuenta. Por un lado, el DENUE presenta información que en principio está actualizada, mientras que los datos con los que se hicieron los agrupamientos iniciales son tomados del censo del año 2010; este desfase de tiempo puede originar algunas incongruencias. Por otra parte, existen muchas agebs que a pesar de no tener determinados tipos de unidades económicas pueden ser igualmente de un nivel socioeconómico alto; por ejemplo, en las unidades habitacionales no suelen existir escuelas u hostipales, sino que estos se encuentran en regiones cercanas. Con este tipo de agebs aparentemente no es posible determinar su nivel socioeconómico mediante las unidades económicas.

7.1. Trabajo a futuro

Como trabajo a futuro, se propone realizar este mismo estudio pero haciendo el agrupamiento inicial con la información del censo de población y vivienda 2020 (cuando

esté disponible). Con esto se puede evitar el problema del desfase de tiempo. También se pueden probar otras formas de manipular las variables en el conjunto de datos de unidades económicas para mejorar la exactitud en la clasificación. En nuestro método tomamos los conjuntos de entrenamiento y prueba de manera aleatorio, lo que puede originar que los resultados no sean los mejores; para ello se puede también aplicar validación cruzada para eliminar las posibles dependencias del modelo con los datos de entrenamiento.

Puede también implementarse algún método geográfico correctivo que asigne la clase correcta a las agebs que no fueron clasificadas correctamente, por ejemplo, si una ageb se encuentra ubicada entre otras dos clasificadas en una clase x, entonces dicha ageb tendrá una alta probabilidad de también pertenecer a la clase x.

Apéndice A

Indicadores propuestos por INEGI

1. Porcentaje de población en viviendas con agua entubada en el ámbito de la vivienda
2. Porcentaje de población en viviendas con energía eléctrica
3. Porcentaje de población en viviendas con drenaje
4. Porcentaje de población en viviendas con piso diferente de tierra
5. Porcentaje de población en viviendas con paredes de materiales durables
6. Porcentaje de población en viviendas con techos de materiales durables
7. Porcentaje de población en viviendas sin hacinamiento
8. Porcentaje de población en viviendas con servicio sanitario exclusivo
9. Porcentaje de población en viviendas que usan gas o electricidad para cocinar
10. Porcentaje de población en viviendas con refrigerador
11. Porcentaje de población en viviendas con radio, radiograbadora o televisión
12. Porcentaje de población en viviendas con teléfono
13. Porcentaje de población en viviendas con automóvil o camioneta propios
14. Porcentaje de población con derechohabiencia a servicios de salud
15. Porcentaje de población de 15 años y más alfabeta
16. Porcentaje de niños de 6 a 14 años que asisten a la escuela
17. Porcentaje de adolescentes de 12 a 17 años que asisten a la escuela

18. Porcentaje de población de 15 años y más con instrucción postprimaria
19. Porcentaje de población ocupada femenina
20. Porcentaje de población económicamente activa entre 20 y 49 años
21. Perceptores por cada 100 personas
22. Porcentaje de población ocupada que percibe más de dos y medio salarios mínimos
23. Porcentaje de población ocupada que percibe más de cinco salarios mínimos
24. Porcentaje de población en hogares que perciben más de $10.42 diarios por persona
25. Porcentaje de población ocupada que son trabajadores familiares sin pago
26. Porcentaje de población ocupada en el sector terciario formal
27. Porcentaje de población ocupada que son profesionistas o técnicos
28. Porcentaje de hijos sobrevivientes de mujeres de 20 a 34 años
29. Segregación de género en términos de alfabetismo
30. Porcentaje de población económicamente inactiva de 65 años y más que es jubilada o pensionada

Apéndice B

Fórmulas para los nuevos indicadores

Indicador	Fórmula
IND01	vph_aguadv / vivpar_hab
IND02	vph_c_elec / vivpar_hab
IND03	vph_pisodt / vivpar_hab
IND04	vph_pisodt / vivpar_hab
IND05	vph_excsa / vivpar_hab
IND06	vph_refri / vivpar_hab
IND07	vph_tv / vivpar_hab
IND08	vph_telef / vivpar_hab
IND09	vph_cel / vivpar_hab
IND10	vph_inter / vivpar_hab
IND11	vph_pc / vivpar_hab
IND12	vph_lavad / vivpar_hab
IND13	vph_autom / vivpar_hab
IND14	pder_ss / pobtot
IND15	(p_15ymas − p15ym_an) / p_15ymas
IND16	(p_6a11 − p6a11_noa) / p_6a11
IND17	(p_12a14 − p12a14noa) / p_12a14
IND18	p15a17a / p_15a17
IND19	p15sec_co / p_15ymas
IND20	p18ym_pb / p_18ymas
IND21	pocupada_f / p_12ymas_f
IND22	pea / p_12ymas

Tabla B.1: Fórmulas para calcular los nuevos indicadores. Los mnemónicos son los mismos que emplea el INEGI en su conjunto de datos del censo del 2010.

Apéndice C

Distribución de agebs por estado acorde al nivel socioeconómico

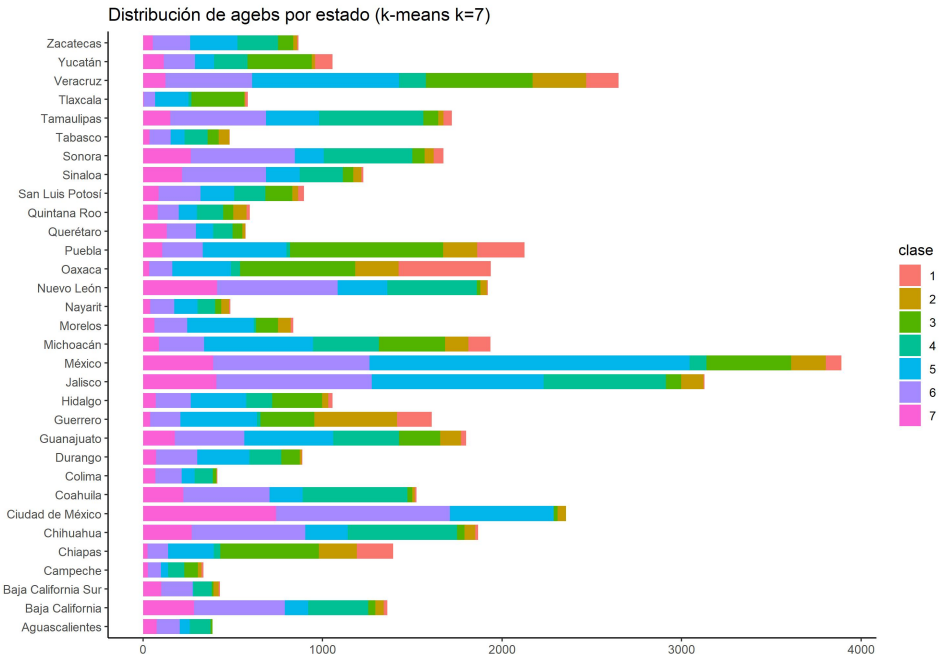

Figura C.1: Distribución de agebs por estado para $k = 7$ (k-means).

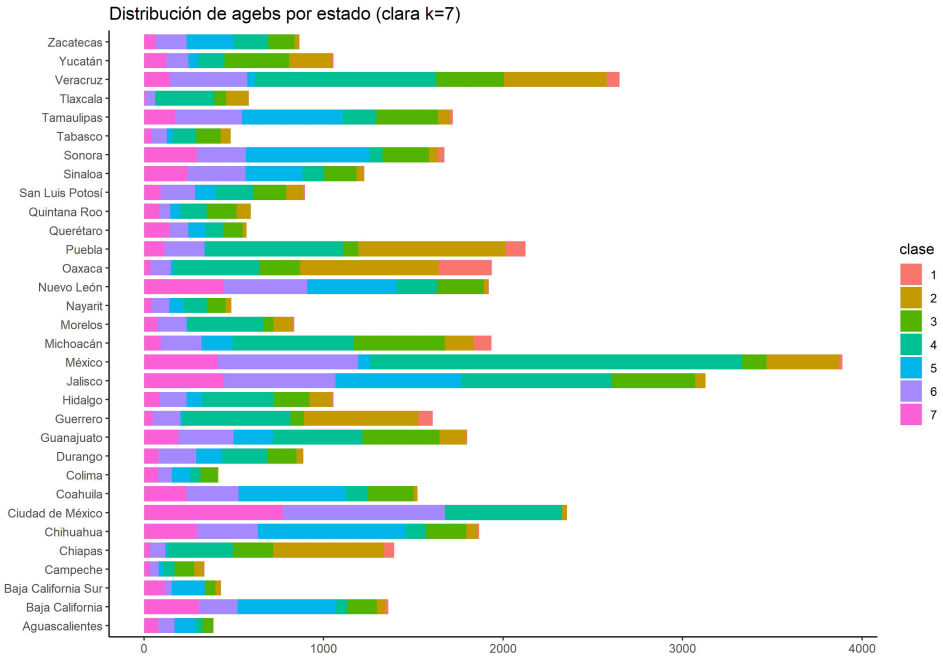

Figura C.2: Distribución de agebs por estado para $k = 7$ (clara).

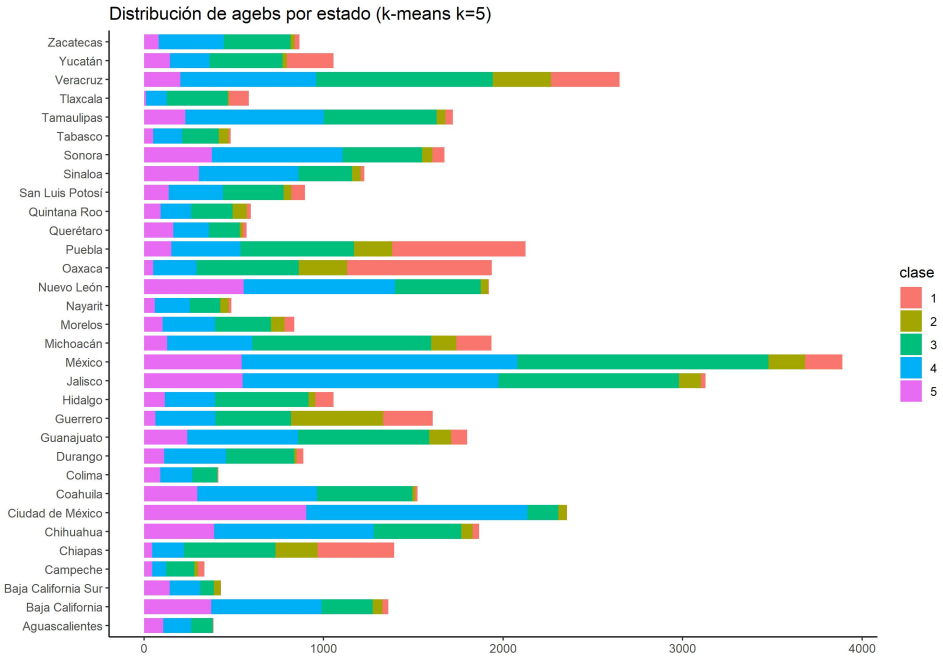

Figura C.3: Distribución de agebs por estado para $k = 5$ (k-means).

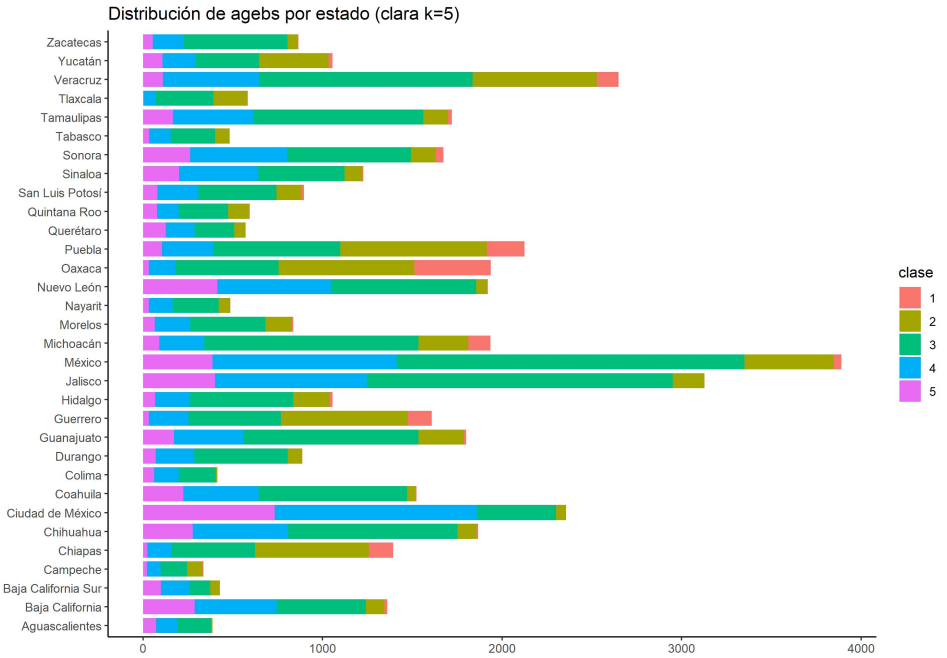

Figura C.4: Distribución de agebs por estado para $k = 5$ (clara).

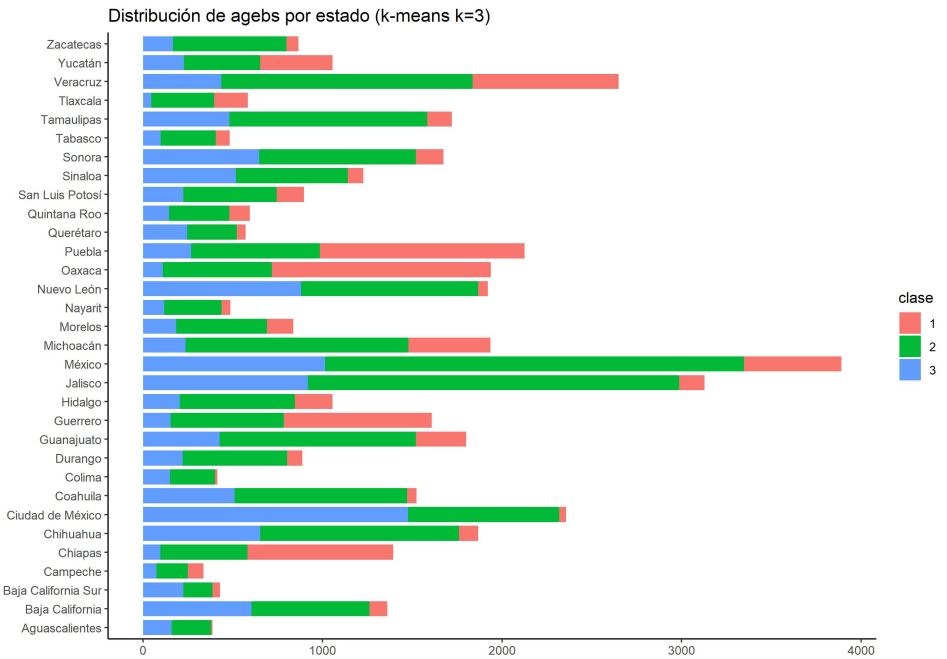

Figura C.5: Distribución de agebs por estado para $k = 3$ (k-means).

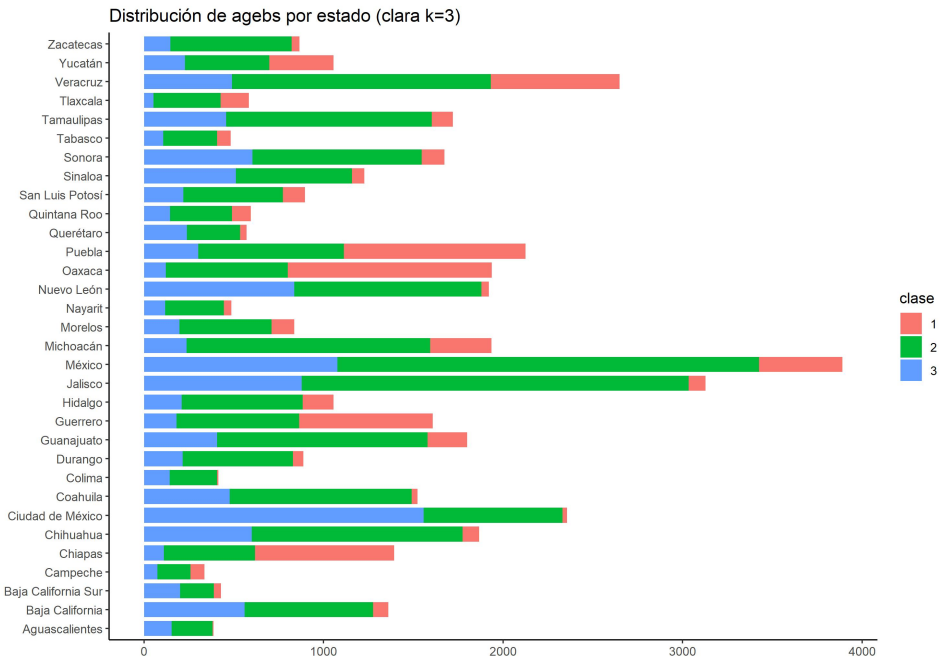

Figura C.6: Distribución de agebs por estado para $k = 3$ (clara).

Apéndice D

Medidas de evaluación para la clasificación de ageb con 5 y 7 clases.

Medidas de evaluación de rf con k-means $k = 5$			
clase	prec	rec	F
1	0.7010870	0.250728863	0.369362921
2	0.3333333	0.001416431	0.002820874
3	0.4764009	0.683663115	0.561516825
4	0.5947830	0.655187541	0.623525728
5	0.7664042	0.447623914	0.565161290

Tabla D.1: Medidas de evaluación para el bosque aleatorio con 5 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rf con clara $k = 5$			
clase	prec	rec	F
1	0.5675676	0.06140351	0.1108179
2	0.4600871	0.17960340	0.2583537
3	0.5755981	0.86546763	0.6913793
4	0.5918544	0.43503185	0.5014684
5	0.7734668	0.44588745	0.5656751

Tabla D.2: Medidas de evaluación para el bosque aleatorio con 5 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de gbm con k-means $k = 5$			
clase	prec	rec	F
1	0.6336842	0.2925170	0.4002660
2	0	0	0
3	0.4751857	0.7141045	0.5706467
4	0.6149452	0.5902610	0.6023503
5	0.7017045	0.5048544	0.5872214

Tabla D.3: Medidas de evaluación para el algoritmo de boosting con 5 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de gbm con clara $k = 5$			
clase	prec	rec	F
1	0.5576923	0.08479532	0.1472081
2	0.4980695	0.14617564	0.2260184
3	0.5688419	0.87014388	0.6879488
4	0.5966746	0.40000000	0.4789323
5	0.6999013	0.51154401	0.5910796

Tabla D.4: Medidas de evaluación para el algoritmo de boosting con 5 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de rna con k-means $k = 5$			
clase	prec	rec	F
1	0.6333333	0.184645287	0.285929270
2	0.2500000	0.002832861	0.005602241
3	0.4682081	0.719178082	0.567170151
4	0.6212121	0.602544418	0.611735887
5	0.7102273	0.510986203	0.594353640

Tabla D.5: Medidas de evaluación para la red neuronal con 5 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rna con clara $k = 5$			
clase	prec	rec	F
1	0.5087719	0.08479532	0.1453634
2	0.5023364	0.12181303	0.1960784
3	0.5715663	0.85323741	0.6845599
4	0.5826011	0.42229299	0.4896603
5	0.6731449	0.54978355	0.6052423

Tabla D.6: Medidas de evaluación para la red neuronal con 5 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de svm con k-means $k = 5$			
clase	prec	rec	F
1	0.7619048	0.1088435	0.1904762
2	0	0	0
3	0.4554169	0.7592593	0.5693361
4	0.6094401	0.5777583	0.5931764
5	0.7282986	0.4287174	0.5397234

Tabla D.7: Medidas de evaluación para la máquina de soporte vectorial con 5 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de svm con clara $k = 5$			
clase	prec	rec	F
1	0.5384615	0.02046784	0.03943662
2	0.5434783	0.04249292	0.07882291
3	0.5437554	0.90521583	0.67940065
4	0.5593870	0.37197452	0.44682479
5	0.7249284	0.36507937	0.48560461

Tabla D.8: Medidas de evaluación para la máquina de soporte vectorial con 5 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de C5.0 con k-means $k = 5$			
clase	prec	rec	F
1	0.5331126	0.31292517	0.39436620
2	0.1145833	0.01558074	0.02743142
3	0.4747414	0.60553019	0.53221851
4	0.5521313	0.60517657	0.57743826
5	0.6307902	0.47317322	0.54072993

Tabla D.9: Medidas de evaluación para el árbol de decisión con 5 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de C5.0 con clara $k = 5$			
clase	prec	rec	F
1	0.5230769	0.0994152	0.1670762
2	0.4063877	0.2090652	0.2760943
3	0.5716529	0.7856115	0.6617680
4	0.5223396	0.4095541	0.4591217
5	0.6141450	0.4949495	0.5481422

Tabla D.10: Medidas de evaluación para el árbol de decisión con 5 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de rf con k-means $k = 7$			
clase	prec	rec	F
1	0.5915493	0.17391304	0.26880000
2	0.1562500	0.00770416	0.01468429
3	0.3517423	0.33312578	0.34218100
4	0.3568731	0.50670241	0.41879016
5	0.4864458	0.52846859	0.50658720
6	0.5009592	0.58477287	0.53963100
7	0.7355278	0.46022727	0.56618611

Tabla D.11: Medidas de evaluación para el bosque aleatorio con 7 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rf con clara $k = 7$			
clase	prec	rec	F
1	0	0	0
2	0.3722628	0.3412639	0.3560900
3	0.3056283	0.3007083	0.3031483
4	0.4768081	0.6430477	0.5475890
5	0.3828475	0.3594737	0.3707926
6	0.5132468	0.4369748	0.4720497
7	0.6873935	0.5230071	0.5940375

Tabla D.12: Medidas de evaluación para el bosque aleatorio con 7 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de gbm con k-means $k = 7$			
clase	prec	rec	F
1	0.6153846	0.18219462	0.28115016
2	0.2083333	0.00770416	0.01485884
3	0.3470919	0.34557908	0.34633385
4	0.3489446	0.56729223	0.43210129
5	0.4803480	0.52388743	0.50117389
6	0.5299663	0.50351887	0.51640420
7	0.6739927	0.52272727	0.58880000

Tabla D.13: Medidas de evaluación para el algoritmo de boosting con 7 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de gbm con clara $k = 7$			
clase	prec	rec	F
1	0.2727273	0.02752294	0.0500000
2	0.3606928	0.35613383	0.3583988
3	0.3076471	0.33676755	0.3215493
4	0.4853911	0.61073229	0.5408954
5	0.3961957	0.38368421	0.3898396
6	0.5260082	0.39805396	0.4531722
7	0.6350148	0.55476345	0.5921826

Tabla D.14: Medidas de evaluación para el algoritmo de boosting con 7 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de rna con k-means $k = 7$			
clase	prec	rec	F
1	0.6093750	0.161490683	0.255319149
2	0.4285714	0.004622496	0.009146341
3	0.3243097	0.372976339	0.346944686
4	0.3539531	0.518498660	0.420709158
5	0.4760758	0.517670157	0.496002508
6	0.5126937	0.497440819	0.504952103
7	0.6651825	0.530539773	0.590280522

Tabla D.15: Medidas de evaluación para la red neuronal con 7 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de rna con clara $k = 7$			
clase	prec	rec	F
1	0.3888889	0.03211009	0.05932203
2	0.3394375	0.38587361	0.36116910
3	0.2913097	0.30650354	0.29871352
4	0.4926058	0.57278387	0.52967786
5	0.3828689	0.39052632	0.38665972
6	0.5010823	0.40955330	0.45071794
7	0.6472393	0.54698639	0.59290481

Tabla D.16: Medidas de evaluación para la red neuronal con 7 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de svm con k-means $k = 7$			
clase	prec	rec	F
1	0.6734694	0.06832298	0.1240602
2	0	0	0
3	0.3457031	0.22042341	0.2692015
4	0.3036981	0.64289544	0.4125237
5	0.4851580	0.49738220	0.4911941
6	0.4911064	0.52111324	0.5056651
7	0.7160665	0.36718750	0.4854460

Tabla D.17: Medidas de evaluación para la máquina de soporte vectorial con 7 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de svm con clara $k = 7$			
clase	prec	rec	F
1	0	0	0
2	0.4640719	0.1152416	0.1846337
3	0.2598471	0.5692209	0.3568113
4	0.4797920	0.6018381	0.5339295
5	0.3942094	0.2794737	0.3270711
6	0.5135301	0.3693056	0.4296373
7	0.6434852	0.5217110	0.5762348

Tabla D.18: Medidas de evaluación para la máquina de soporte vectorial con 7 clases. Agrupamiento inicial hecho con clara.

Medidas de evaluación de C5.0 con k-means $k = 7$			
clase	prec	rec	F
1	0.4085603	0.21739130	0.28378378
2	0.1822660	0.05701079	0.08685446
3	0.3040201	0.30136986	0.30268918
4	0.3418269	0.38123324	0.36045627
5	0.4132863	0.46007853	0.43542893
6	0.4389810	0.47408829	0.45585974
7	0.5455963	0.49715909	0.52025269

Tabla D.19: Medidas de evaluación para el árbol de decisión con 7 clases. Agrupamiento inicial hecho con k-means.

Medidas de evaluación de C5.0 con clara $k = 7$			
clase	prec	rec	F
1	0.1764706	0.02752294	0.04761905
2	0.3526448	0.31226766	0.33123028
3	0.2763550	0.29877656	0.28712871
4	0.4550717	0.53602135	0.49224068
5	0.3407407	0.31473684	0.32722298
6	0.4143673	0.38522778	0.39926656
7	0.5558602	0.52559948	0.54030646

Tabla D.20: Medidas de evaluación para el árbol de decisión con 7 clases. Agrupamiento inicial hecho con clara.

Bibliografía

[AMAI, 2018] AMAI (2018). Nivel Socio Económico AMAI 2018. Technical report, Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión.

[Berzofsky et al., 2007] Berzofsky, M., Smiley-McDonald, H., Moore, A., and Krebs, C. (2007). Measuring Socioeconomic Status (SES) in the NCVS: Background, Options, and Recommendations. Technical report, Bureau of Justice Statistics.

[Boser et al., 1992] Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

[Bradley and Corwyn, 2002] Bradley, R. H. and Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, 53(1):371–399. PMID: 11752490.

[Breunig et al., 2000] Breunig, M. M., Kriegel, H.-P., Ng, R., and Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. volume 29, pages 93–104.

[Commons, 2012] Commons, W. (2012). Svm separating hyperplanes.

[Commons, 2015] Commons, W. (2015). A neural network with multiple layers.

[Downey, 2011] Downey, A. B. (2011). *Think Stats: Probability and Statistics for Programmers*. O'Reilly Media, 1 edition.

[Galobardes et al., 2006] Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., and Davey Smith, G. (2006). Indicators of socioeconomic position (part 1). *Journal of Epidemiology & Community Health*, 60(1):7–12.

[Gao et al., 2011] Gao, J., Hu, W., Zhang, Z., Zhang, X., and Wu, O. (2011). RKOF: Robust Kernel-Based Local Outlier Detection. volume 6635, pages 270–283.

[Harris et al., 2005] Harris, R., Sleight, P., and Webber, R. (2005). *Geodemographics, GIS and Neighbourhood Targeting*. Wiley.

[Hastie et al., 2009] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2 edition.

[Hsu et al., 2003] Hsu, C.-W., Chang, C.-C., and Lin, C.-J. (2003). A Practical Guide to Support Vector Classification. https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf. Accessed: 2019-08-20.

[INEGI, a] INEGI. Censo de Población y Vivienda 2010. https://www.inegi.org.mx/programas/ccpv/2010/. Accessed: 2019-08-26.

[INEGI, b] INEGI. Clasificadores - Catálogo SCIAN. https://www.inegi.org.mx/app/scian/. Accessed: 2019-08-26.

[INEGI, c] INEGI. Directorio Estadístico Nacional de Unidades Económicas. https://www.inegi.org.mx/app/mapa/denue/. Accessed: 2019-08-26.

[INEGI, d] INEGI. Regiones Socioeconómicas de México. http://sc.inegi.gob.mx/niveles/index.jsp. Accessed: 2019-06-05.

[INEGI, 2004] INEGI (2004). Regiones Socioeconómicas de México. Technical report, Instituto Nacioal de Estadística, Geografía e Informática.

[James et al., 2013] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer, 1st edition.

[Jarque, 1981] Jarque, C. M. (1981). A solution to the problem of optimum stratification in multivariate sampling. *The Journal of the royal statistical society*, 30(2):163.

[Kaufman and Rousseeuw, 2009] Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

[Larose and Larose, 2014] Larose, D. T. and Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley, 2 edition.

[Liberatos et al., 1988] Liberatos, P., Link, B. G., and Kelsey, J. L. (1988). The Measurement of Social Class in Epidemology. *Epidemiologic Reviews*, 10(1):87–121.

[Lynch and Kaplan, 2007] Lynch, J. and Kaplan, G. (2007). Socioeconomic position. *Social Epidemiology*.

[McLaren, 2007] McLaren, L. (2007). Socioeconomic Status and Obesity. *Epidemiologic Reviews*, 29(1):29–48.

[Mohri et al., 2012] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 1 edition.

[Murphy, 2012] Murphy, K. (2012). *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning series. MIT Press, 1 edition.

[Oakes and Rossi, 2003] Oakes, M. J. and Rossi, P. H. (2003). The measurement of ses in health research: Current practice and steps toward a new approach. *Social science & medicine (1982)*, 56:769–84.

[Ripley, 1996] Ripley, B. D. (1996). *Patern Recognition and Neural Networks*. Cambridge University Press.

[Shavers, 2007] Shavers, V. L. (2007). Measurement of socioeconomic status in health disparities research. *Journal of the National Medical Association*, 99:1013–1023.

[Tan et al., 2013] Tan, P.-N., Steinbach, M., and Kumar, V. (2013). *Introduction to Data Mining*. Pearson, 2 edition.

[Werner et al., 2007] Werner, S., Malaspina, D., and Rabinowitz, J. (2007). Socioeconomic Status at Birth Is Associated With Risk of Schizophrenia: Population-Based Multilevel Study. *Schizophrenia Bulletin*, 33(6):1373–1378.

[Zaki et al., 2014] Zaki, M. J., Meira Jr, W., and Meira, W. (2014). *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press.